International Journal of Computer Vision
https://doi.org/10.1007/s11263-020-01350-x

Semantically Tied Paired Cycle Consistency for Any-Shot Sketch-Based Image Retrieval

Anjan Dutta[1] · **Zeynep Akata**[2]

Received: 14 May 2019 / Accepted: 19 June 2020

Abstract

Low-shot sketch-based image retrieval is an emerging task in computer vision, allowing to retrieve natural images relevant to hand-drawn sketch queries that are rarely seen during the training phase. Related prior works either require aligned sketch-image pairs that are costly to obtain or inefficient memory fusion layer for mapping the visual information to a semantic space. In this paper, we address any-shot, *i.e.* zero-shot and few-shot, sketch-based image retrieval (SBIR) tasks, where we introduce the few-shot setting for SBIR. For solving these tasks, we propose a semantically aligned paired cycle-consistent generative adversarial network (SEM-PCYC) for any-shot SBIR, where each branch of the generative adversarial network maps the visual information from sketch and image to a common semantic space via adversarial training. Each of these branches maintains cycle consistency that only requires supervision at the category level, and avoids the need of aligned sketch-image pairs. A classification criteria on the generators' outputs ensures the visual to semantic space mapping to be class-specific. Furthermore, we propose to combine textual and hierarchical side information via an auto-encoder that selects discriminating side information within a same end-to-end model. Our results demonstrate a significant boost in any-shot SBIR performance over the state-of-the-art on the extended version of the challenging Sketchy, TU-Berlin and QuickDraw datasets.

1 Introduction

Matching natural images with free-hand sketches, *i.e. sketch-based image retrieval* (SBIR) (Yu et al. 2015, 2016a; Liu et al. 2017; Pang et al. 2017; Song et al. 2017b; Shen et al. 2018; Zhang et al. 2018; Chen and Fang 2018; Kiran Yelamarthi et al. 2018; Dutta and Akata 2019; Dey et al. 2019) has received a lot of attention. Since sketches can effectively express shape, pose and some 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 may be because in some situations it is difficult 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 on a touch screen.

As the visual information from all classes gets explored by the system during training, with overlapping training and test classes, existing SBIR methods perform well (Zhang et al. 2018). Since for practical applications there is no guarantee that the training data would include all possible queries, a more realistic setting is *low-shot* or *any-shot* SBIR (AS-SBIR) (Shen et al. 2018; Kiran Yelamarthi et al. 2018; Dutta and Akata 2019; Dey et al. 2019), which combines zero- and few-shot learning (Lampert et al. 2014; Vinyals et al. 2016; Xian et al. 2018a; Ravi and Larochelle 2017) and SBIR as a single task, where the aim is an accurate class prediction and a competent retrieval performance. However, this is an extremely challenging task, as it simultaneously deals with domain gap, intra-class variability and limited or no knowledge on *novel* classes. Additionally, fine-grained SBIR (Pang et al. 2017, 2019) is an alternative sketch-based image retrieval task, allowing to search for specific object images,

Communicated by Jun-Yan Zhu, Hongsheng Li, Eli Shechtman, Ming-Yu Liu, Jan Kautz, Antonio Torralba.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11263-020-01350-x) contains supplementary material, which is available to authorized users.

✉ Anjan Dutta
a.dutta@exeter.ac.uk

Zeynep Akata
zeynep.akata@uni-tuebingen.de

[1] Department of Computer Science, Innovation Centre, University of Exeter, Streatham Campus, Exeter EX4 4RN, UK

[2] Cluster of Excellence Machine Learning, Tübingen AI Center, University of Tübingen, 72076 Tübingen, Germany

Published online: 29 July 2020

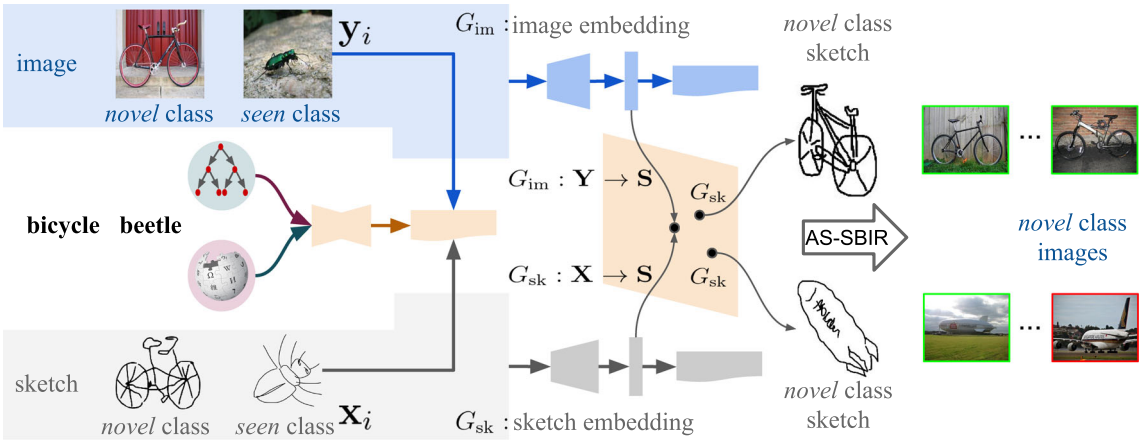

Fig. 1 Our SEM-PCYC model learns to map visual information from *seen*-class sketches and images to a semantic space through an adversarial training procedure in zero-shot SBIR setting. Furthermore, our mdoel is flexible enough to use a few examples from *novel* classes to fine-tune the model, where the *novel* classes contain a few labeled samples in few-shot SBIR (FS-SBIR) setting. During the testing phase the learned mappings are used to generate embeddings of the *novel* classes. We refer to the combination of zero- and few-shot SBIR as any-shot SBIR (AS-SBIR)

which has already received remarkable attention in the computer vision community. However, it has never been explored in low shot setting, which is an extremely challenging and at the same time of high practical relevance.

One of the major shortcomings of the prior work on any-shot SBIR is that a natural image is retrieved after learning a mapping from an input sketch to an output image using a training set of labelled *aligned* pairs (Kiran Yelamarthi et al. 2018). The supervision of the pair correspondence is to enhance the correlation of multi-modal data (here, sketch and image) so that learning can be guided by semantics. However, for many realistic scenarios, paired (aligned) training data is either unavailable or obtaining it is very expensive. Furthermore, often a joint representation of two or more modalities is learned by using a memory fusion layer (Shen et al. 2018), such as, tensor fusion (Hu et al. 2017), bilinear pooling (Yu et al. 2017) etc. These fusion layers are often expensive in terms of memory (Yu et al. 2017), and extracting useful information from this high dimensional space could result in information loss (Yu et al. 2018).

To alleviate these shortcomings, we propose a semantically aligned paired cycle consistent generative adversarial network (SEM-PCYC) model for any-shot SBIR task, where each branch either maps the 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 that the mapping of sketch or image modality to a common semantic space and their translation back to the original modality, avoiding 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 any-shot SBIR. Furthermore, inspired by the previous works on label embedding (Akata et al. 2015), we propose to combine side information from text-based and hierarchical models via a feature selection auto-encoder (Wang et al. 2017) which selects discriminating side information based on intra and inter class covariance.

This paper extends our CVPR 2019 conference paper (Dutta and Akata 2019), with the following additional contributions: (1) We propose to apply the SEM-PCYC model for any-shot SBIR task, *i.e.* addition to zero-shot paradigm, we introduce few-shot setting for SBIR and combine it with generalized setting, which has been experimentally proven to be effective for difficult or confusing classes (Fig. 1). (2) We adapt the recent zero-shot SBIR models and ours to fine-grained SBIR in the generalized low-shot setting and provide an extensive benchmark including quantitative and qualitative evaluations. (3) We evaluate our model on one recent dataset, i.e. QuickDraw, in addition to extending our experiments to new settings with Sketchy and TU-Berlin. We show that our proposed model consistently improves the state-of-the-art results of any-shot SBIR on all the three datasets.

2 Related Work

As our work belongs at the verge of sketch-based image retrieval and any-shot learning task, we briefly review the relevant literature from these 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 (Liu et al. 2017). 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 (Hu and Collomosse 2013), histogram of oriented edges (Saavedra 2014), learned key shapes (Saavedra and Barrios 2015) 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 (Qi et al. 2016; Sangkloy et al. 2016; Yu et al. 2016a; Song et al. 2017a) that are suitable for cross-modal retrieval. These frameworks either use generic ranking losses, *viz.*, contrastive loss (Chopra et al. 2005), triplet ranking loss (Sangkloy et al. 2016) or more sophisticated HOLEF based loss (Song et al. 2017b) for the same. Further to these discriminative losses, Pang et al. (2017) introduced a discriminative-generative hybrid model for preserving all the domain invariant information useful for reducing the domain gap between sketch and image. Alternatively, Liu et al. (2017) and Zhang et al. (2018) focus on learning cross-modal hash code for category level SBIR within an end-to-end deep model.

In addition to the above coarse-grained SBIR models, fine-grained sketch-based image retrieval (FG-SBIR) has gained popularity recently (Li et al. 2014; Song et al. 2017a, b; Pang et al. 2017). In this more realistic setting, a FG-SBIR model allows to search a specific object or image. First, models tackled this task using deformable part model and graph matching (Li et al. 2014). Recently, different ranking frameworks and corresponding losses, such as, siamese (Pang et al. 2017), triplet (Sangkloy et al. 2016), quadruplet (Song et al. 2017a) networks were used for the same. Song et al. (2017b) proposed attention model for FG-SBIR task, Zhang et al. (2018) improving retrieval efficiency using a hashing scheme.

Zero-Shot Learning (ZSL) and Few-Shot Learning (FSL). 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 Xian et al. (2018a). Early works on ZSL (Lampert et al. 2014; Jayaraman and Grauman 2014; Changpinyo et al. 2016; Al-Halah et al. 2016) 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 (Frome et al. 2013; Romera-Paredes and Torr 2015; Akata et al. 2015, 2016; Kodirov et al. 2017) directly learn a mapping from image feature space to a semantic space. Many other ZSL approaches learn non-linear multi-modal embedding (Socher et al. 2013; Akata et al. 2016; Xian et al. 2016; Changpinyo et al. 2017; Zhang et al. 2017), 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 (Zhang and Saligrama 2015; Fu et al. 2015; Zhang and Saligrama 2016; Akata et al. 2016; Long et al. 2017). Although, most of the deep neural network models in this domain are trained using a discriminative loss function, a few generative models also exist (Wang et al. 2018a; Xian et al. 2018b; Chen et al. 2018) 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 (Lampert et al. 2014) that, however, require costly expert annotation. Thus, there has been a large group of studies (Mensink et al. 2014; Akata et al. 2015; Xian et al. 2016; Reed et al. 2016; Qiao et al. 2016; Ding et al. 2017) which utilize other auxiliary information, such as, text-based (Mikolov et al. 2013) or hierarchical model (Miller 1995) for label embedding.

On the other hand, few-shot learning (FSL) refers to the task of recognizing images or detecting objects with a model trained on very few samples (Xian et al. 2019; Schönfeld et al. 2018). Directly training a given model with small amount of training samples could have the risk of over fitting. Hence a general step to overcome this hurdle is to initially train the model on classes with sufficient examples, and then generalize it to classes with fewer examples without learning any new parameters. This setup already attracted a lot of attention within the computer vision community. One of the first attempts (Koch et al. 2015) is a siamese convolutional network model for computing similarity between pair of images, and then the learned similarity was used to solve the one-shot problem by k-nearest neighbors classification. On the other hand, matching network model (Vinyals et al. 2016) uses cosine distance to predict image label based on support sets and apply the episodic training strategy that mimics few-shot learning. An extension, i.e. prototypical network (Snell et al. 2017), used Euclidean distance instead of cosine distance and built a prototype representation of each class for the few-shot learning scenario. As an orthogonal direction Ravi and Larochelle (2017) introduced meta-learning framework for FSL, which updates weights of a classifier for a given episode. Model agnostic meta-learner (Finn et al. 2017) learns better weight initialization capable to generalize in FSL scenario with fewer gradient descent steps. There also exist few low shot methods that learn a generator from the base class data to generate novel class features for data augmentation (Girshick 2015; Wang et al. 2018b). Alternatively, GNN (Kipf and Welling 2017) was also proposed as a

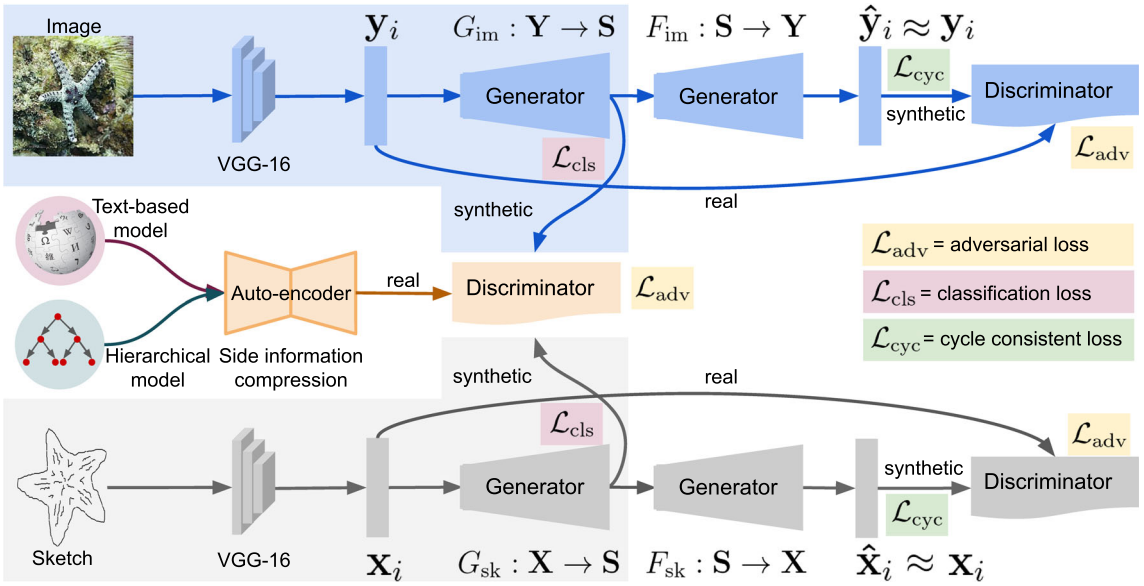

Fig. 2 Our SEM-PCYC Model. The sketch (in light gray) and image cycle consistent networks (in light blue) respectively map the sketch and image to the semantic space and then the original input space. An auto-encoder (light orange) combines the semantic information based on text and hierarchical model, 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 polygonal region) and image (light blue polygonal region) encoders to the semantic space are used for generating embeddings on the *novel* classes for any-shot, *i.e.* zero- and few-shot SBIR. (best viewed in color) (color figure online)

framework for few-shot learning task (Satorras and Estrach 2018).

Our Work. The prior work on zero-shot sketch-based image retrieval (ZS-SBIR) (Shen et al. 2018), proposed a generative cross-modal hashing scheme using a graph convolution network for aligning the sketch and image in the semantic space. Inspired by them, Kiran Yelamarthi et al. (2018) 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 this work, 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. The cycle consistency constraint on each branch allows supervision only at category level, and avoids the need of aligned sketch-image pairs. Furthermore, we address zero-shot and few-shot cross-modal (sketch to image) retrieval, for that, we effectively combine different side information within an end-to-end framework, and map visual information to the semantic space through an adversarial training. Finally, we unify low-shot learning models and generalize them to fine-grained SBIR scenario.

3 Semantically Aligned Paired Cycle Consistent GAN (SEM-PCYC)

Our Semantically Aligned Paired Cycle Consistent GAN (SEM-PCYC) model uses the sketch and image data from the *seen* categories for training the underlying model. It then encodes and matches the sketch and image categories that remain novel during the training phase. The overall pipeline of our end-to-end deep architecture is shown in Fig. 2.

We define $\mathcal{D}^s = \{\mathbf{X}^s, \mathbf{Y}^s\}$ to be a collection of sketch-image data from the training categories $\mathcal{C}^s$, which contains sketch images $\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^N$ as well as natural images $\mathbf{Y}^s = \{\mathbf{y}_i^s\}_{i=1}^N$, where N is the total number of sketch and image pairs that are not necessarily aligned. Without loss of generality, a sketch and an image have the same index i, and share the same category label. The set $\mathbf{S}^s = \{\mathbf{s}_i^s\}_{i=1}^N$ indicates the side information necessary for transferring knowledge from seen to the novel classes (a.k.a unseen classes in zero-shot learning literature). In our setting, we also use an auxiliary training set $\mathcal{D}^a = \{\mathbf{X}^a, \mathbf{Y}^a\}$ from the unseen classes $\mathcal{C}^u$ which is disjoint from $\mathcal{C}^s$, where the number of samples per class is fixed to k.

Our aim 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 is applied to the novel classes. Now, given a second set $\mathcal{D}^u = \{\mathbf{X}^u, \mathbf{Y}^u\}$ from the test categories $\mathcal{C}^u$, the proposed deep networks

$G_{\text{sk}} : \mathbb{R}^d \rightarrow \mathbb{R}^M$, $G_{\text{im}} : \mathbb{R}^d \rightarrow \mathbb{R}^M$ (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. Depending on k, *i.e.* the number of samples considered per class as an auxiliary set, the scenario is called k-shot. In the classical zero-shot sketch-based image retrieval setting, the test categories belong to $\mathcal{C}^u$, in other words, at test time the assumption is that every image will come from a previously unseen class. This is not realistic as the true generalization performance of the classifier can only be measured with how well it generalizes to unseen classes without forgetting the classes it has seen. Hence, in the generalized zero-shot sketch based image retrieval scenario the search space contains both $\mathcal{C}^u$ and $\mathcal{C}^s$. In other words, at test time an image may come either from a previously seen or an unseen class. As this setting is significantly more challenging, the accuracy decreases for all the methods considered.

3.1 Paired Cycle Consistent Generative Model

To achieve the flexibility to handle sketch and image individually, *i.e.* even without 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 the 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. (2017) 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 $G_{\text{sk}} : \mathbf{X} \rightarrow \mathbf{S}$ 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 $G_{\text{im}} : \mathbf{Y} \rightarrow \mathbf{S}$ can be trained via the same discriminator such that $\hat{\mathbf{s}}_i = G_{\text{im}}(\mathbf{y}_i)$.

Adversarial Loss. As shown in Fig. 2, for mapping the sketch and image representation to a common semantic space, we introduce four generators $G_{\text{sk}} : \mathbf{X} \rightarrow \mathbf{S}$, $G_{\text{im}} : \mathbf{Y} \rightarrow \mathbf{S}$, $F_{\text{sk}} : \mathbf{S} \rightarrow \mathbf{X}$ and $F_{\text{im}} : \mathbf{S} \rightarrow \mathbf{Y}$. 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:

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

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, namely

$$\min_{G_{\text{sk}}, G_{\text{im}}} \max_{D_{\text{se}}} \mathcal{L}_{\text{adv}}(G_{\text{sk}}, G_{\text{im}}, D_{\text{se}}, \mathbf{x}, \mathbf{y}, \mathbf{s}).$$

In a similar way, for the generator F_{sk} and its discriminator D_{sk}, the objective is:

$$\begin{aligned}\mathcal{L}_{\text{adv}}(F_{\text{sk}}, D_{\text{sk}}, \mathbf{x}, \mathbf{s}) = \mathbb{E}\left[\log D_{\text{sk}}(\mathbf{x})\right] \\ + \mathbb{E}\left[\log(1 - D_{\text{sk}}(F_{\text{sk}}(\mathbf{s})))\right]\end{aligned} \tag{2}$$

F_{sk} minimizes the objective and its adversary D_{sk} intends to maximize it, namely

$$\min_{F_{\text{sk}}} \max_{D_{\text{sk}}} \mathcal{L}_{\text{adv}}(F_{\text{sk}}, D_{\text{sk}}, \mathbf{x}, \mathbf{s}).$$

Similarly, another adversarial loss is introduced for the mapping F_{im} and its discriminator D_{im}, *i.e.* $\min_{F_{\text{im}}} \max_{D_{\text{im}}} \mathcal{L}_{\text{adv}}(F_{\text{im}}, D_{\text{im}}, \mathbf{y}, \mathbf{s})$.

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 (Zhu et al. 2017). 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 $G_{\text{sk}} : \mathbf{X} \rightarrow \mathbf{S}$ and another mapping $F_{\text{sk}} : \mathbf{S} \rightarrow \mathbf{X}$, then both G_{sk} and F_{sk} are reverse of each other, and hence form a one-to-one correspondence or bijective mapping.

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

where $\mathbf{s}$ is the semantic features of the class c which is the category label of $\mathbf{x}$. Similarly, a cycle consistency loss is imposed for the mappings $G_{\text{im}} : \mathbf{Y} \rightarrow \mathbf{S}$ and $F_{\text{im}} : \mathbf{S} \rightarrow \mathbf{Y}$: $\mathcal{L}_{\text{cyc}}(G_{\text{im}}, F_{\text{im}})$. 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.

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

where c is the category label of $\mathbf{x}$, $P(c|G_{\text{sk}}(\mathbf{x}); \theta)$ denotes the probability of $G_{\text{sk}}(\mathbf{x})$ being predicted with its true class label c. The conditional probability is computed by a linear softmax classifier parameterized by θ. Similarly, a classification loss $\mathcal{L}_{\text{cls}}(G_{\text{im}})$ is also imposed on the generator G_{im}.

3.2 Selection of Side Information

Learning a compatibility or a matching function between multiple modalities in zero-shot scenario (Shen et al. 2018; Dey et al. 2019; Liu et al. 2019) requires structure in the class embedding space where the image features are mapped to. Attributes provide one such a structured class embedding space (Lampert et al. 2014), however obtaining attributes requires costly human annotation. On the other hand, side information can also be learned at a much lower cost from large-scale text corpora such as Wikipedia. Similarly, output embeddings built from hierarchical organization of classes such as WordNet can also provide structure in the output space and substitute the attributes. Motivated by attribute selection for zero-shot learning (Guo et al. 2018), 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:

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

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. Additionally, $\|.\|_F$ denotes the Frobenius norm defined as the square root of the sum of the absolute squares of its elements and $\|.\|_{2,1}$ indicates $\ell_{2,1}$ norm (Nie et al. 2010). Selecting side information reduces the dimensionality of embeddings, which further improves retrieval time. Therefore, the training objective of our model:

$$\begin{aligned}
&\mathcal{L}(G_{\text{sk}}, G_{\text{im}}, F_{\text{sk}}, F_{\text{im}}, D_{\text{se}}, D_{\text{sk}}, D_{\text{im}}, f, g, \mathbf{x}, \mathbf{y}, \mathbf{s}) \\
&= \lambda^{\text{se}}_{\text{adv}}\mathcal{L}_{\text{adv}}(G_{\text{sk}}, G_{\text{im}}, D_{\text{se}}, \mathbf{x}, \mathbf{y}, \mathbf{s}) \\
&\quad + \lambda^{\text{sk}}_{\text{adv}}\mathcal{L}_{\text{adv}}(F_{\text{sk}}, D_{\text{sk}}, \mathbf{x}, \mathbf{s}) \\
&\quad + \lambda^{\text{im}}_{\text{adv}}\mathcal{L}_{\text{adv}}(F_{\text{im}}, D_{\text{im}}, \mathbf{y}, \mathbf{s}) \\
&\quad + \lambda^{\text{sk}}_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{\text{sk}}, F_{\text{sk}}) + \lambda^{\text{im}}_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{\text{im}}, F_{\text{im}}) \\
&\quad + \lambda^{\text{sk}}_{\text{cls}}\mathcal{L}_{\text{cls}}(G_{\text{sk}}) + \lambda^{\text{im}}_{\text{cls}}\mathcal{L}_{\text{cls}}(G_{\text{im}}) \\
&\quad + \lambda_{\text{aenc}}\mathcal{L}_{\text{aenc}}(f, g)
\end{aligned} \tag{6}$$

where different λs are the weights on respective loss terms. For obtaining the initial side information, we combine a text-based and a hierarchical model, which are complementary and robust (Akata et al. 2015). Below, we provide a description of our text-based and hierarchical models for side information.

Text-based Model. We use three different text-based side information. (1) Word2Vec (Mikolov et al. 2013) 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 (Pennington et al. 2014) 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. (3) FastText (Joulin et al. 2017) extends the Word2Vec model, where instead of learning vector for words directly, FastText represents each word as n-gram of characters and then trains a skip-gram model to learn the embeddings. FastText works well with rare words, even if a word was not seen during training, it can be broken down into n-grams to get its embeddings, which is a huge advantage of this model.

Hierarchical Model. Semantic distance (or similarity) between words can also be approximated by their distance (or similarity) in a large ontology such as WordNet[1] with

[1] https://wordnet.princeton.edu.

$\approx$ 100, 000 words in English. One can measure the similarity [$\mathcal{S}_{\text{WN}}$ in Eq. (7)] between words represented as nodes in the ontology using techniques, such as *path similarity*, e.g. counting the number of hops required to reach from one node to the other, and Jiang and Conrath (1997). For a set $\mathbb{S}$ of nodes in a dictionary $\mathbb{D}$ that consists of a set of classes, similarities between every class c and all the other nodes considered in the same order in $\mathbb{S}$ to determine the entries of the class embedding vector (Akata et al. 2015) of c [$\mathbf{s}_{\text{hier}}(c)$ in Eq. (7)]:

$$\mathbf{s}_{\text{hier}}(c) = [\mathcal{S}_{\text{WN}}(c, c_1), \ldots, \mathcal{S}_{\text{WN}}(c, c_{|\mathbb{S}|})] \tag{7}$$

Note that, $\mathbb{S}$ considers all the nodes on the path from each node in $\mathbb{D}$ to its highest level ancestor. The WordNet hierarchy contains most of the classes of the Sketchy (Sangkloy et al. 2016), Tu-Berlin (Eitz et al. 2012) and QuickDraw (Dey et al. 2019) 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. $|\mathbb{S}|$, i.e. the number of nodes, for Sketchy, TU-Berlin and QuickDraw datasets are respectively 354, 664 and 344.

4 Experiments

In this section, we detail our datasets, implementation protocol and present our results on (generalized) zero-shot, (generalized) few-shot and fine-grained settings.

Datasets. We experimentally validate our model on three popular SBIR datasets, namely Sketchy (Extended), TU-Berlin (Extended) and QuickDraw (Extended). For brevity, we refer to these extended datasets as Sketchy, TU-Berlin and QuickDraw respectively.

The Sketchy Dataset (Sangkloy et al. 2016) is a large collection of sketch-photo pairs. The dataset originally consists of images from 125 different classes, with 100 photos each. The 75,471 sketch images of the objects that appear in these 12,500 images are collected via crowd sourcing. This dataset also contains a fine grained correspondence (alignment) between particular photos and sketches as well as various data augmentations for deep learning based methods. Liu et al. (2017) extended the dataset by adding 60,502 photos yielding in total 73,002 images. We randomly pick 25 classes as the *novel* test set, and the data from remaining 100 *training* classes.

The original TU-Berlin Dataset (Eitz et al. 2012) contains 250 categories with a total of 20,000 sketches extended by Liu et al. (2017) with 204,489 natural images corresponding to the sketch classes. 30 classes of sketches and images are randomly chosen to respectively from the query set and the retrieval gallery. The remaining 220 classes are utilized for training. We follow Shen et al. (2018) and select classes with at least 400 images to form a test set.

The QuickDraw (Extended), a large-scale dataset proposed recently in Dey et al. (2019), contains the sketch-image pairs of 110 classes consisting of 203,885 images and 330,111 sketches, i.e. approximately 1854 images/class and 3000 sketches/class. The main difference of this dataset from the previous ones is in the abstractness of the sketches which are collected from the *Quick, Draw!*[2] online game. The increased abstractness in the drawings has eventually enlarged the sketch-image domain gap, and hence increased the challenge of SBIR task.

Implementation details. We implemented the SEM-PCYC model using PyTorch (Paszke et al. 2017) deep learning toolbox[3] on a single TITAN Xp or TITAN V graphics card. Unless otherwise mentioned, we extract features from sketch and image from the VGG-16 (Simonyan and Zisserman 2014) network model pre-trained on ImageNet (Deng et al. 2009) (before the last pooling layer). In Sect. 4.1, we compare the VGG-16 features with SE-ResNet-50 features for zero-shot SBIR task, which is only restricted to that experimentation. 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. (2017b) without the shortcut connection for extracting only the informative regions from sketch and image. The attended 512d 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, 220 classes for TU-Berlin and 80 classes for QuickDraw). Both the generators G_{sk} and G_{im} are built with a fc layer followed by a ReLU non-linearity that accept 512d vector and output Md representation, whereas, the generators F_{sk} and F_{im} take Md features and produce 512d 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. We experimentally set $\lambda^{\text{se}}_{\text{adv}} = 1.0$, $\lambda^{\text{sk}}_{\text{adv}} = 0.5$, $\lambda^{\text{im}}_{\text{adv}} = 0.5$, $\lambda^{\text{sk}}_{\text{cyc}} = 1.0$, $\lambda^{\text{im}}_{\text{cyc}} = 1.0$, $\lambda^{\text{sk}}_{\text{cls}} = 1.0$, $\lambda^{\text{im}}_{\text{cls}} = 1.0$, $\lambda_{\text{aenc}} = 0.01$ to give the optimum performance of our model.

While constructing the hierarchy for the class embedding, we only consider the *training* classes belong to that dataset. In this way, the WordNet hierarchy or the knowledge graph for the Sketchy, TU-Berlin and QuickDraw datasets respectively

[2] https://quickdraw.withgoogle.com.

[3] Our code and models are available at: https://github.com/AnjanDutta/sempcyc-ijcv.

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 (Yang et al. 2016a), ZSIH (Shen et al. 2018), GDH (Zhang et al. 2018), that produce hash codes, we have used the iterative quantization (ITQ) (Gong et al. 2013) algorithm to obtain the binary 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 (Generalized) Zero-Shot Sketch-Based Image Retrieval

Apart from the two prior Zero-Shot SBIR works closest to ours, *i.e.* ZSIH (Shen et al. 2018) and ZS-SBIR (Kiran Yelamarthi et al. 2018), we adopt fourteen ZSL and SBIR models to the zero-shot SBIR task. Note that in this setting, the training classes are indicated as "seen" and novel classes as "unseen" since none of the sketches of these classes are visible to the model during training.

The SBIR methods that we evaluate are SaN (Yu et al. 2015), 3D Shape (Wang et al. 2015a), Siamese CNN (Qi et al. 2016), GN Triplet (Sangkloy et al. 2016), DSH (Liu et al. 2017) and GDH (Zhang et al. 2018). A softmax baseline is also added, which is based on computing the 4096d VGG-16 (Simonyan and Zisserman 2014) feature vector pretrained on the *seen* classes for nearest neighbour search. The ZSL methods that we evaluate are: CMT (Socher et al. 2013), DeViSE (Frome et al. 2013), SSE (Zhang and Saligrama 2015), JLSE (Zhang and Saligrama 2016), ZSH (Yang et al. 2016a), SAE (Kodirov et al. 2017) and FRWGAN (Felix et al. 2018). 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) (Su et al. 2015; Shen et al. 2018) 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 between the sketch and image modalities. The majority SBIR methods although have performed better than their ZSL counterparts, fail to generalize the learned representations to *unseen* classes. However, GN Triplet (Sangkloy et al. 2016), DSH (Liu et al. 2017), GDH (Zhang et al. 2018) 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 or QuickDraw datasets, as in these datasets, aligned sketch-image pairs are not available. Our proposed method has excels the state-of-the-art method by 0.091 mAP@all on the Sketchy, 0.074 mAP@all on the TU-Berlin and 0.046 mAP@all on the QuickDraw, which shows the effectiveness of our proposed SEM-PCYC model due to the cycle consistency between sketch, image and semantic space, as well as the compact and discriminative side information.

In general, the main challenge in TU-Berlin dataset is the large number of visually similar and overlapping classes. On the other hand, in QuickDraw datatset there is a the large domain gap that is intentionally introduced for designing future realistic models. Also, the ambiguity in annotation, e.g. non-professional sketches, is a major challenge in this dataset. Although our 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, we hope that our work will encourage further research in improving these results.

Finally, the PR-curves of SEM-PCYC and considered baselines on Sketchy, TU-Berlin and QuickDraw are respectively shown in Fig. 3a–c which show that the precision-recall curves correspond to our SEM-PCYC model (dark blue line) are always plotted above the other methods. This indicates that our proposed model consistently exhibits the superiority on all three datasets, which clearly show the benefit of our proposal.

Generalized Zero-Shot Sketch-Based Image Retrieval. We conducted experiments on generalized ZS-SBIR setting where search space contains both *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 show that our model significantly outperforms both the existing models (Shen et al. 2018; Kiran Yelamarthi et al. 2018), due to the benefit of our cross-modal adversarial mechanism and heterogeneous side information.

Qualitative Results. We analyze the retrieval performance of our proposed model qualitatively in Figs. 4, 5 and 6. Some notable examples are as follows. Sketch query of `tank` retrieves some examples of `motorcycle` probably because both of them have wheels in common (row 1 of Fig. 4). Similar explanation can be given in the case of `car` and `motorcycle` (row 1 of Fig. 6). For having visual and semantic similarity, sketching `guitar` retrieves some `violins` (row 2 of Fig. 4). This can also be observed in case of `train` and `van` in row 2 of Fig. 6.

For having visual and semantic similarity, querying `bear` retrieves some `squirrels` (row 3 of Fig. 4). Querying objects with wheel (*e.g.*, `wheelchair`, `motorcycle`)

Table 1 (Generalized) zero-shot sketch-based image retrieval and (generalized) fine-grained sketch-based image retrieval performance comparison with existing SBIR, ZSL, zero-shot SBIR and generalized zero-shot SBIR methods

	Method	Sketchy (extended)				TU-Berlin (extended)				QuickDraw (extended)			
		mAP @all	Prec. @100	Feat. dim	Retr. Time (s)	mAP @all	Prec. @100	Feat. dim	Retr. Time (s)	mAP @all	Prec. @100	Feat. dim	Retr. Time (s)
SBIR	Softmax Baseline	0.114	0.172	4096	3.5×10^{-1}	0.089	0.143	4096	4.3×10^{-1}	0.058	0.095	4096	4.6×10^{-1}
	Siamese CNN (Qi et al. 2016)	0.132	0.175	64	5.7×10^{-3}	0.109	0.141	64	5.9×10^{-3}	0.074	0.112	64	5.8×10^{-3}
	SaN (Yu et al. 2016b)	0.115	0.125	512	4.8×10^{-2}	0.089	0.108	512	5.5×10^{-2}	0.060	0.093	512	5.9×10^{-2}
	GN Triplet (Sangkloy et al. 2016)	0.204	0.296	1024	9.1×10^{-2}	0.175	0.253	1024	1.9×10^{-1}	0.118	0.142	1024	2.3×10^{-1}
	3D Shape (Wang et al. 2015b)	0.067	0.078	64	7.8×10^{-3}	0.054	0.067	64	7.2×10^{-3}	0.036	0.081	64	8.1×10^{-1}
	DSH (binary) (Liu et al. 2017)	0.171	0.231	64	6.1×10^{-5}	0.129	0.189	64	7.2×10^{-5}	0.087	0.127	64	7.6×10^{-5}
	GDH (binary) (Zhang et al. 2018)	0.187	0.259	64	7.8×10^{-5}	0.135	0.212	64	9.6×10^{-5}	0.095	0.146	64	1.1×10^{-4}
ZSL	CMT (Socher et al. 2013)	0.087	0.102	300	2.8×10^{-2}	0.062	0.078	300	3.3×10^{-2}	0.036	0.062	300	3.6×10^{-2}
	DeViSE (Frome et al. 2013)	0.067	0.077	300	3.6×10^{-2}	0.059	0.071	300	3.2×10^{-2}	0.034	0.073	300	3.4×10^{-2}
	SSE (Zhang et al. 2015)	0.116	0.161	100	1.3×10^{-2}	0.089	0.121	220	1.7×10^{-2}	0.051	0.093	80	1.8×10^{-2}
	JLSE (Zhang and Saligrama 2016)	0.131	0.185	100	1.5×10^{-2}	0.109	0.155	220	1.4×10^{-2}	0.063	0.084	80	1.5×10^{-2}
	SAE (Kodirov et al. 2017)	0.216	0.293	300	2.9×10^{-2}	0.167	0.221	300	3.2×10^{-2}	0.096	0.112	300	3.3×10^{-2}
	FRWGAN (Felix et al. 2018)	0.127	0.169	512	3.2×10^{-2}	0.110	0.157	512	3.9×10^{-2}	0.064	0.093	512	4.2×10^{-2}
	ZSH (binary) (Yang et al. 2016b)	0.159	0.214	64	5.9×10^{-5}	0.141	0.177	64	7.6×10^{-5}	0.081	0.118	64	7.8×10^{-5}
Zero-shot SBIR	ZSIH (binary) (Shen et al. 2018)	0.258	0.342	64	6.7×10^{-5}	0.223	0.294	64	7.7×10^{-5}	0.131	0.188	64	7.9×10^{-5}
	ZS-SBIR (Kiran Yelamarthi et al. 2018)	0.196	0.284	1024	9.6×10^{-2}	0.005	0.001	1024	1.2×10^{-1}	0.006	0.001	1024	1.6×10^{-1}
	SEM-PCYC	**0.349**	**0.463**	64	1.7×10^{-3}	**0.297**	**0.426**	64	1.9×10^{-3}	**0.177**	**0.255**	64	2.1×10^{-3}
	SEM-PCYC (binary)	**0.344**	**0.399**	64	9.5×10^{-5}	**0.293**	**0.392**	64	9.3×10^{-4}	**0.164**	**0.243**	64	9.6×10^{-4}
Generalized zero-shot SBIR	ZSIH (binary) (Shen et al. 2018)	0.219	0.296	64	6.7×10^{-5}	0.142	0.218	64	7.7×10^{-5}	0.130	0.163	64	8.1×10^{-5}
	ZS-SBIR (Kiran Yelamarthi et al. 2018)	0.146	0.190	1024	7.8×10^{-2}	0.003	0.001	1024	6.7×10^{-2}	0.002	0.001	1024	8.2×10^{-2}
	SEM-PCYC	**0.307**	**0.364**	64	1.7×10^{-3}	**0.192**	**0.298**	64	2.0×10^{-3}	**0.140**	**0.221**	64	2.1×10^{-4}
	SEM-PCYC (binary)	**0.260**	**0.317**	64	9.4×10^{-5}	**0.174**	**0.267**	64	9.3×10^{-4}	**0.135**	**0.216**	64	9.4×10^{-4}

SBIR and ZSL methods are adapted to the zero-shot SBIR task, same *seen* and *unseen* classes are used for a fair comparison.
Best results are highlighted in bold

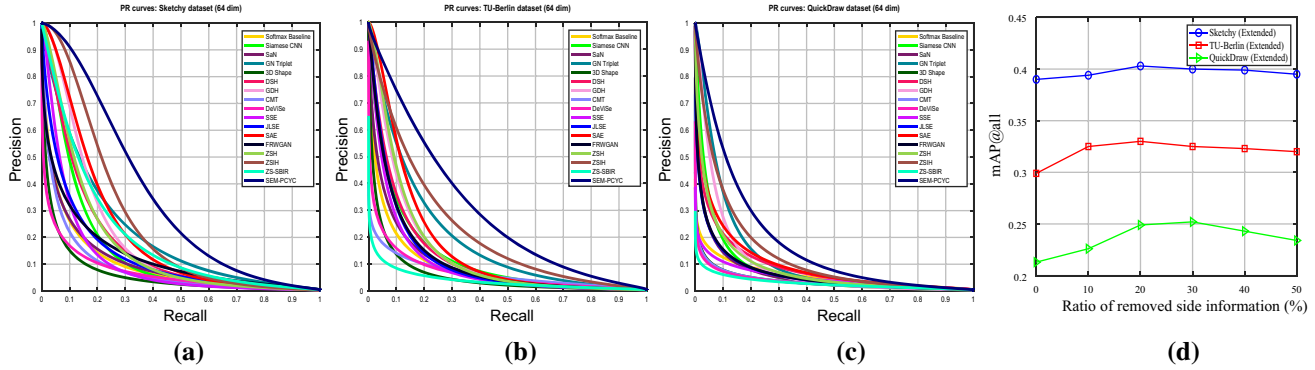

Fig. 3 **a**–**c** PR curves of SEM-PCYC model and several SBIR, ZSL and zero-shot SBIR methods respectively on the Sketchy, TU-Berlin and QuickDraw datasets, **d** plot showing mAP@all wrt the ratio of removed side information. (best viewed in color) (color figure online)

Fig. 4 Top-20 zero-shot SBIR results obtained by our SEM-PCYC model on the Sketchy (extended) dataset are shown here according to the Euclidean distances, where the green ticks denote the correctly retrieved candidates, whereas the red crosses indicate the wrong retrievals. (best viewed in color) (color figure online)

sometime wrongly retrieves other vehicles, probably because of having wheels in common (row 6 of Fig. 4). Sketch query of `spoon` retrieves some examples of `racket` (row 4 of Fig. 4), possibly for having significant visual similarity. Sketch of `burger` retrieves some examples of `jack-o-lantern` (row 5 of Fig. 4), perhaps for having same shape. Querying `castle`, retrieves images having large portion of sky (row 2 of Fig. 5), because the images of its semantically similar classes, such as, `skyscraper`, `church`, are mostly captured with sky in background. Similar phenomenon can be observed in case of `tree` and `electrical post` in row 5 of Fig. 6. Querying `duck`, retrieves images of `swan` or `shark` (row 4 of Fig. 5), probably for having watery background in common. Sketch of `pickup truck` retrieves some images from `traffic light` class for having a truck like object in the scene (row 3 of Fig. 5). Sketching `bookshelf` retrieves some examples of `cabinet` for having significant visual and semantic similarity (row 5 of Fig. 5).

Sometimes too much abstraction in sketches can produce wrong retrieval results. For example, in row 3 of Fig. 6, it is difficult to understand whether the sketch is of `eiffel tower` or any other tower or a hill. Furthermore, we have observed certain ambiguities in annotation of images in QuickDraw dataset. Currently, the images are much complex, which often contain two or more objects, and most of the currently available SBIR datasets provide single object annotation ignoring the object in background. For example see row 6 of Fig. 6, many of the wrongly retrieved images truly contain `flower`, whereas some of them are annotated

Fig. 5 Top-20 zero-shot SBIR results obtained by our SEM-PCYC model on the TU-Berlin (extended) dataset are shown here according to the Euclidean distances, where the green ticks denote the correctly retrieved candidates, whereas the red crosses indicate the wrong retrievals. (best viewed in color) (color figure online)

Fig. 6 Top-20 zero-shot SBIR results obtained by our SEM-PCYC model on the QuickDraw (extended) dataset are shown here according to the Euclidean distances, where the green ticks denote the correctly retrieved candidates, whereas the red crosses indicate the wrong retrievals. (best viewed in color) (color figure online)

as `tower` or `trees` etc. Additionally, as the images from QuickDraw dataset are collected from the Flickr website, it contains many subsequent captures which can be confused as identical frames. Hence, although some retrievals on QuickDraw dataset appear identical, they are not in terms of the actual pixel values.

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 Fig. 7, the classes `swan`, `duck` and `owl`, `penguin` have substantial visual similarity, and all of them are `standing bird`

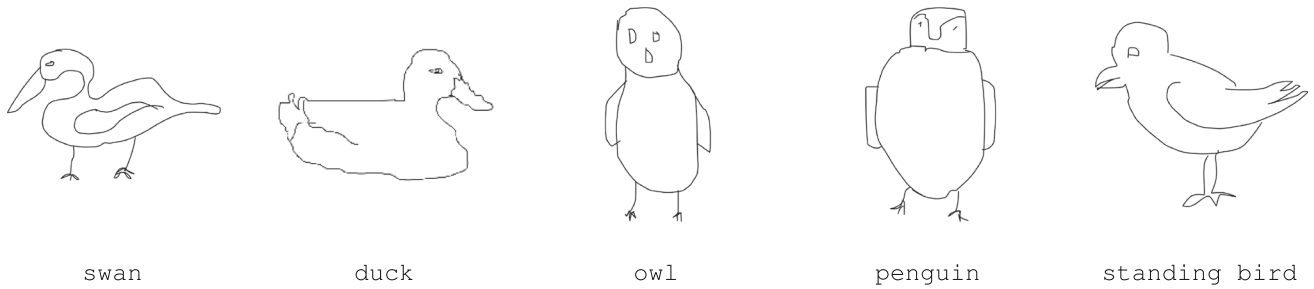

Fig. 7 Inter-class similarity in TU-Berlin dataset may indicate the challenge of the task

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.

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 (Pennington et al. 2014), Word2Vec (Mikolov et al. 2013) and FastText (Joulin et al. 2017) as text-based model, and three similarity measurements, *i.e.* path, Lin (1998) and Jiang-Conrath (Jiang and Conrath 1997) for constructing three different side information that are based on WordNet hierarchy. Table 2 contains the quantitative results on Sketchy, TU-Berlin and QuickDraw datasets with different side information mentioned and their combinations, where we set $M = 32, 64, 128$. We have observed that in majority of cases combining different side information increases the performance by 1–3%.

On Sketchy, the combination of Word2Vec and Jiang-Conrath hierarchical similarity as well as FastText and Path reach the highest mAP of 0.349 with 64d embedding while on TU Berlin dataset, in addition to the combination of Word2Vec and path similarity, FastText and Path lead with 0.297 mAP with 64d, and for QuickDraw the combination of GloVe and Lin hierarchical similarity reaches to 0.177 for 64d. We conclude from these experiments that indeed text-based and hierarchy-based class embeddings are complementary.

Effect of Visual Features. Visual features are also crucial for the zero-shot SBIR task. For having some overview on that, addition to VGG-16 (Simonyan and Zisserman 2014) features obtained before the last `fc` layer, we also consider SE-ResNet-50 (Hu et al. 2019; He et al. 2015) features, and perform zero-shot SBIR experiments on the Sketchy, TU-Berlin and QuickDraw datasets with different semantic models mentioned above. In Table 3, we present the mAP@all values obtained by the considered visual features and semantic models, where we observe that SE-ResNet-50 features work consistently better than VGG-16 on all the three datasets. Especially, the performance gain on the challenging TU-Berlin dataset should be noted, which we speculate as the benefit of feature calibration strategy involved in the SE blocks, that effectively produces robust features minimizing inter-class confusion as presented in Fig. 7.

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 our model by only withdrawing the adversarial loss for the semantic domain, which should indicate the effect of side information in our case. We also 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 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 4. We consider the best side information setting according to Table 2 depending on the dataset. The assessed baselines have typically underperformed 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 (Zhu et al. 2017), but does not guarantee an individual input and output are paired up.

Table 2 Zero-shot SBIR mAP@all using different semantic embeddings (top) and their combinations (bottom) with 32, 64 and 128 dimension

Text embedding			Hierarchical embedding			Sketchy (extended)			TU-Berlin (extended)			QuickDraw (extended)		
GloVe (Pennington et al. 2014)	Word2Vec (Mikolov et al. 2013)	FastText (Joulin et al. 2017)	Path	Lin (Lin 1998)	Ji-Cn (Jiang and Conrath 1997)	32 dim	64 dim	128 dim	32 dim	64 dim	128 dim	32 dim	64 dim	128 dim
✓						0.237	0.284	0.321	0.193	0.228	0.239	0.127	0.149	0.165
	✓					0.279	0.330	0.365	0.199	0.232	0.243	0.124	0.132	0.167
		✓				0.264	0.344	0.343	0.219	0.262	0.265	0.127	0.155	0.165
			✓			0.290	0.314	0.365	0.201	0.224	0.255	0.121	0.138	0.155
				✓		0.201	0.248	0.264	0.152	0.169	0.182	0.130	0.149	0.158
					✓	0.263	0.308	0.352	0.208	0.227	0.239	0.151	0.146	0.152
✓			✓			0.259	0.338	0.356	0.238	0.276	0.281	0.129	0.176	0.158
✓				✓		0.275	0.299	0.318	0.241	0.253	0.264	0.130	**0.177**	0.175
✓					✓	0.273	0.285	0.291	0.238	0.243	0.251	**0.149**	0.163	0.165
	✓		✓			0.298	0.340	0.368	**0.278**	**0.297**	**0.301**	0.145	0.150	0.164
	✓			✓		0.282	0.288	0.306	0.253	0.264	0.282	0.142	0.169	0.175
	✓				✓	**0.307**	**0.349**	0.372	0.273	0.291	0.298	0.145	0.155	**0.184**
		✓	✓			0.329	**0.349**	**0.400**	0.242	**0.297**	0.289	0.137	0.151	0.153
		✓		✓		0.304	0.344	0.352	0.254	0.296	0.286	0.129	0.150	0.147
		✓			✓	0.317	0.299	0.381	0.246	0.279	0.326	0.124	0.144	0.182

Best results are highlighted in bold

Table 3 Zero-shot SBIR mAP@all using different semantic embeddings either with VGG-16 or ResNet-50 visual features while the dimension is kept equal to 64

Visual features	Semantic model	Sketchy (extended)	TU-Berlin (extended)	QuickDraw (extended)
VGG-16 (Simonyan and Zisserman 2014)	GloVe (Pennington et al. 2014)	0.284	0.228	0.149
	Word2Vec (Mikolov et al. 2013)	0.330	0.232	0.132
	FastText (Joulin et al. 2017)	**0.344**	**0.262**	**0.155**
	Path	0.314	0.224	0.138
	Lin (Lin 1998)	0.248	0.169	0.149
	Ji-Cn (Jiang and Conrath 1997)	0.308	0.227	0.146
SE-ResNet-50 (Hu et al. 2019; He et al. 2015)	GloVe (Pennington et al. 2014)	0.344	0.329	**0.172**
	Word2Vec (Mikolov et al. 2013)	**0.385**	0.305	0.151
	FastText (Joulin et al. 2017)	0.368	**0.349**	0.171
	Path	0.330	0.317	0.156
	Lin (Lin 1998)	0.362	0.318	0.146
	Ji-Cn (Jiang and Conrath 1997)	0.384	0.306	0.161

Best results are highlighted in bold

Table 4 Ablation study on our SEM-PCYC model (64d) on three datasets (measured with mAP@all)

Description	Sketchy (extended)	TU-Berlin (extended)	QuickDraw (extended)
Only adversarial loss	0.128	0.109	0.065
Adversarial + cycle consistency loss	0.147	0.131	0.078
Adversarial + classification loss	0.140	0.127	0.076
Adversarial (sketch + image) + cycle consistency + classification loss	0.213	0.154	0.075
Without selecting side information	0.382	0.299	0.185
Without regularizer in Eq. (5)	0.323	0.273	0.158
SEM-PCYC (full model)	**0.349**	**0.297**	**0.177**

Best results are highlighted in bold

Imposing 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 notice that without the adversarial loss for the semantic domain, our model performs better than the previously mentioned three configurations but does not reach near to the full model. This is due to the fact that without semantic mapping, the resulting embeddings are not semantically related to each other, which do not help in cross modal retrieval in zero-shot scenario. We further observe that without the encoded and compact side information, we achieve better mAP@all with a compromise on retrieval time, as the original dimension ($354 + 300 = 654$d for Sketchy, $664 + 300 = 964$d for TU-Berlin and $344 + 300 = 644$d for QuickDraw) 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 Fig. 3c), and we have observed that at the beginning, reducing a small part (mostly 5–30%) 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 20–40%) deteriorates the performance of the system, which is quite justifiable because the compressing mechanism of auto-encoder progressively removes impor-

tant 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.

Finally, without using the regularizer in Eq. (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.

4.2 (Generalized) Few-Shot Sketch-Based Image Retrieval

For the few-shot scenario, we start with the pre-trained model trained in the zero-shot setting, and then fine tune it using a few example images, e.g. k-shot, from "novel" classes. For fine tuning the model in k-shot setting, we consider k different sketch and image instances from each of the *unseen* classes and cross-combine according to the coarse-grained and fine-grained settings to fine tune the model. The performance is evaluated on the rest of the instances from each class at test time.

Few-Shot Sketch-Based Image Retrieval. Figure 8a–c present the few-shot SBIR performance of our SEM-PCYC model together with ZSIH (Shen et al. 2018) and ZS-SBIR (Kiran Yelamarthi et al. 2018) respectively on the Sketchy, TU-Berlin and QuickDraw databases. All these plots show that the considered methods have performed consistently with the increment of k. However, this growth slowly gets saturated after $k = 10$. In this case also our proposed SEM-PCYC model consistently outperforms the other prior works, which clearly points out the supremacy of our proposal.

Generalized Few-Shot Sketch-Based Image Retrieval. We also tested our few-shot model in generalized scenario, where during the test phase the search space includes both the *seen* and *novel* classes. Typically, this setting poses remarkably challenging scenario as the *seen* classes may create significant confusion to the *novel* queries. However, the generalized setting is more realistic as it allows to query the system with sketch from any classes. In this setting as well, we considered ZSIH (Shen et al. 2018) and ZS-SBIR (Kiran Yelamarthi et al. 2018) as two benchmark methods and trained them with the same experimental settings as ours. In FS-SBIR the generalized setting results follow the non-generalized setting quite closely (see Fig. 8d–f). This eventually indicates the convergence of the generalization ability of different models. In this setting as well, our proposed model steadily surpassed both the benchmark models, which indicates the advantage of our proposed model.

Qualitative Results. Figures 9, 10 and 11 present a selection of qualitative results obtained by our SEM-PCYC model respectively on the Sketchy, TU-Berlin and QuickDraw datasets in the scenario of increasing number of shots, which show an evolution of model performance with the increment of k ($= 0, 1, 5, 10$) for the classes where 0-shot results are weak. From these results, we can see that sometimes a single unseen example is sufficient to correctly retrieve images (row 3 of Fig. 9, row 5 of Fig. 10 and row 5 of Fig. 11), however, sometimes it needs more examples (row 2 and 5 of Fig. 9, row 2, 3, 4 of Fig. 10 and row 2, 3, 4 of Fig. 11) to remove the confusion from the other similar classes. This uncertainty may either come from visual or semantic similarity. As expected, increasing the number of examples also improves the performance.

Model Ablations. Similar to zero-shot setting, we perform an ablation study for few-shot scenario as well, where we consider the same model baselines as of Table 4. The mAP@all values obtained by those baselines in 5-shot scenario are shown in Table 5. In this case, all the baselines have achieved much better performance than the corresponding zero-shot performance on that dataset, which is absolutely justified since the model is already trained to zero-shot setting and having few examples from novel classes provide some gain with any combination of losses. We observe that the first three configurations (first three rows of Table 5) work quite closely across all the three datasets and we haven't found any prominent difference among these three baselines on the considered datasets. However, the baselines with more criterion or losses (bottom three rows of Table 5) achieve much better performance from the previously mentioned three baselines. Among these baselines, we have not found much difference between the ones that do and do not use side information. This is due to the consideration of pre-trained zero-shot model which already has past knowledge of side information, and in this case training with side information could be slightly redundant.

Fine-Grained Settings. We have further evaluated our model in fine-grained setting where the task is to find a specific object image of a drawn sketch, and we have combined it with the above mentioned variations of k-shot scenarios. For this experiment, we only considered the Sketchy dataset as only this corpus contains aligned sketch-image pairs, which are often used for fine-grained SBIR evaluation tasks. We have not considered other fine-grained datasets, such as *shoe*, *chair* etc (Song et al. 2017a) as they do not contain class information which we need for semantic space mapping. For this setting as well, we have considered ZSIH (Shen et al. 2018) and ZS-SBIR (Kiran Yelamarthi et al. 2018) as the two benchmark methods and the same experimental protocol.

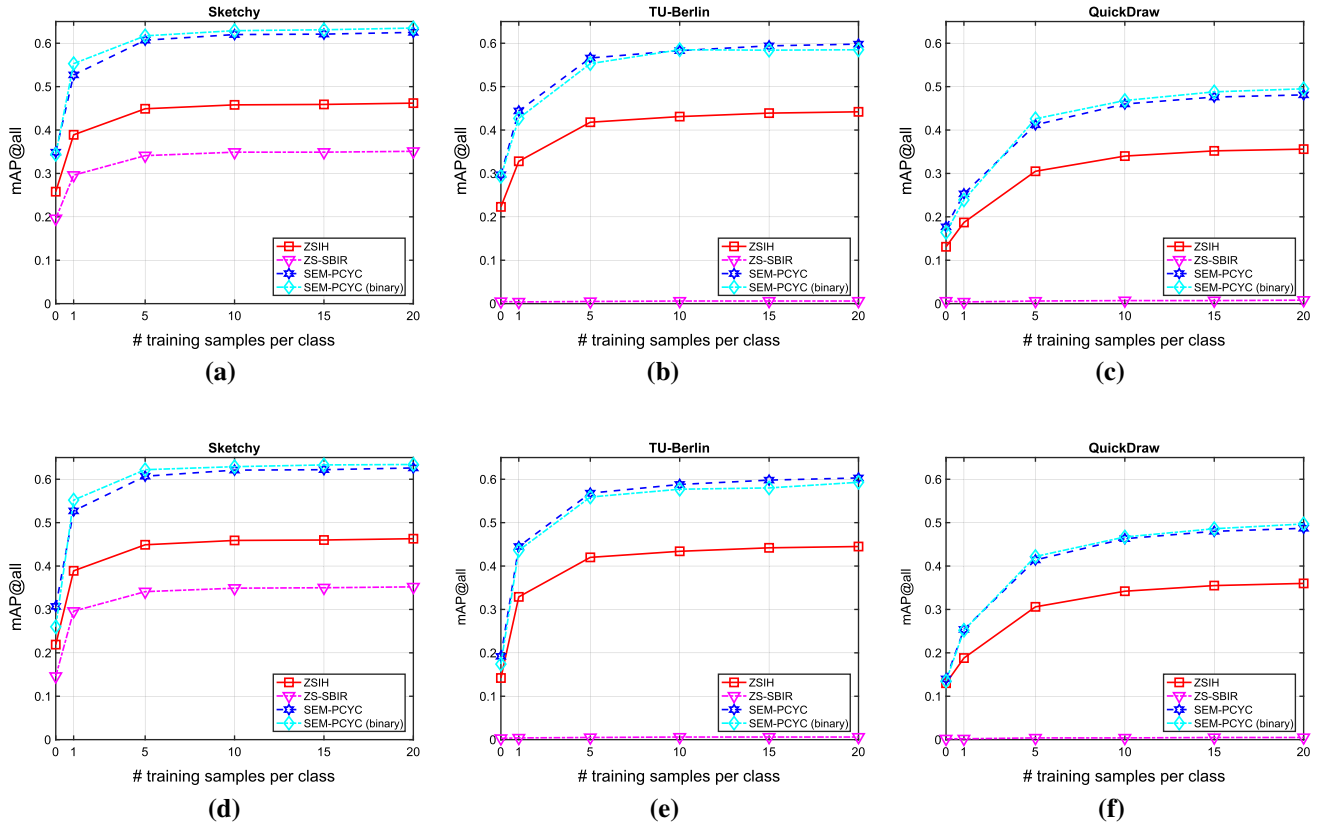

Fig. 8 Few-shot sketch-based image retrieval (k = 0, 1, 5, 10, 15, 20) performance comparison with three existing state-of-the-art methods on Sketchy, TU-Berlin and Quickdraw datasets. Top: few-shot sketch based image retrieval results, Bottom: generalized few-shot sketch-based image retrieval results (color figure online)

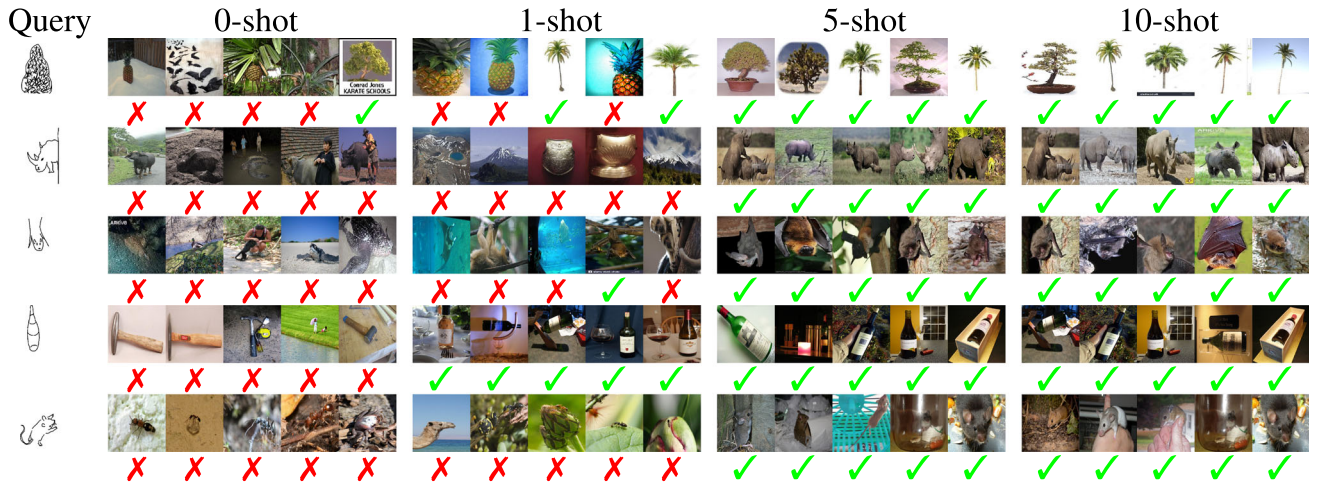

Fig. 9 Top-5 k-shot ($k = 0, 1, 5, 10$) SBIR results obtained by our SEM-PCYC model on the Sketchy (extended) dataset are shown here according to the Euclidean distances, where the green ticks denote the correctly retrieved candidates, whereas the red crosses indicate the wrong retrievals. (best viewed in color) (color figure online)

Figure 12a and b show the performance of our model in fine-grained generalized few-shot together with ZSIH (Shen et al. 2018) and ZS-SBIR (Kiran Yelamarthi et al. 2018). In fine-grained setting, all the methods have performed remarkably poor. We explain this fact as the drawback of semantic space mapping which intends to map visual information from

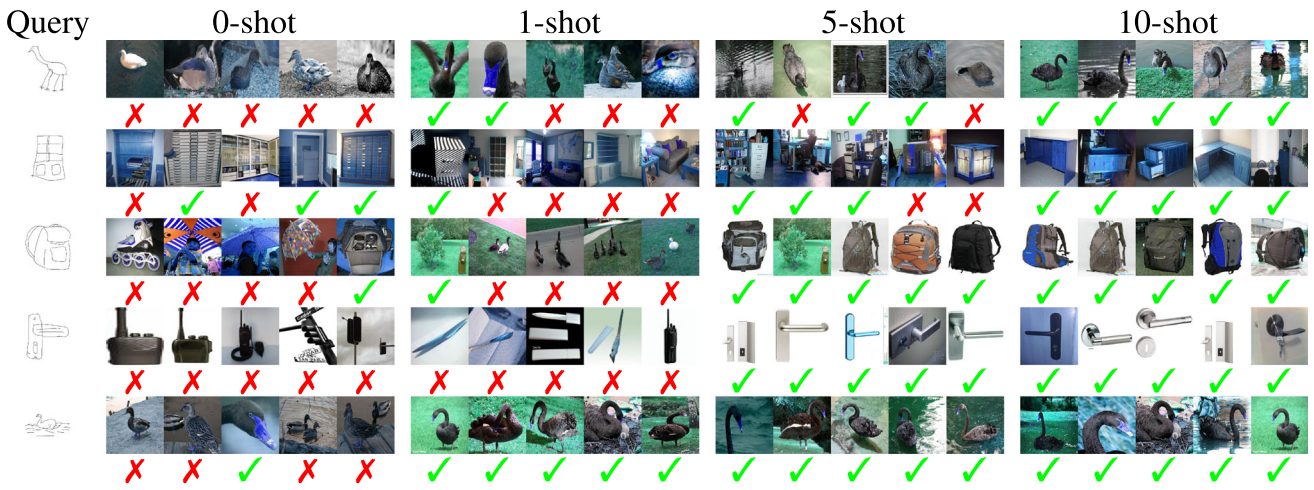

Fig. 10 Top-5 k-shot ($k = 0, 1, 5, 10$) SBIR results obtained by our SEM-PCYC model on the TU-Berlin (extended) dataset are shown here according to the Euclidean distances, where the green ticks denote the correctly retrieved candidates, whereas the red crosses indicate the wrong retrievals. (best viewed in color) (color figure online)

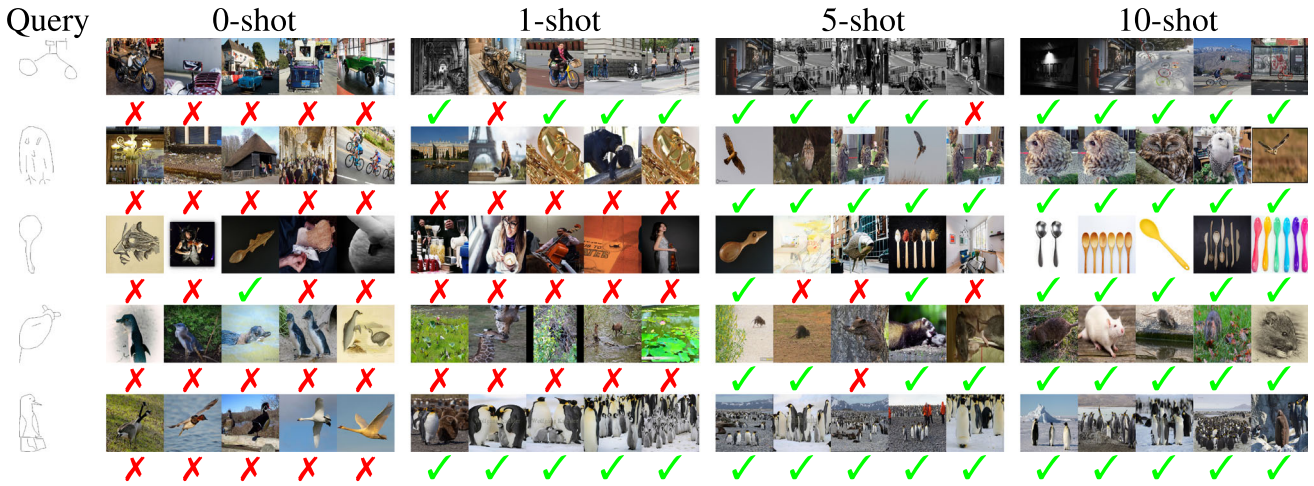

Fig. 11 Top-5 k-shot ($k = 0, 1, 5, 10$) SBIR results obtained by our SEM-PCYC model on the QuickDraw (extended) dataset are shown here according to the Euclidean distances, where the green ticks denote the correctly retrieved candidates, whereas the red crosses indicate the wrong retrievals. (best viewed in color) (color figure online)

Table 5 Ablation study with *few shot* setting on our SEM-PCYC model (64d) on three datasets (measured with mAP@all)

Description	Sketchy (5-shot)	TU-Berlin (5-shot)	QuickDraw (5-shot)
Only adversarial loss	0.512	0.489	0.312
Adversarial + cycle consistency loss	0.508	0.499	0.307
Adversarial + classification loss	0.534	0.483	0.298
Adversarial (sketch + image) + cycle consistency + classification loss	0.592	0.559	0.378
Without regularizer in Eq. (5)	0.602	0.543	0.365
SEM-PCYC (full model)	**0.607**	**0.566**	**0.412**

Best results are highlighted in bold

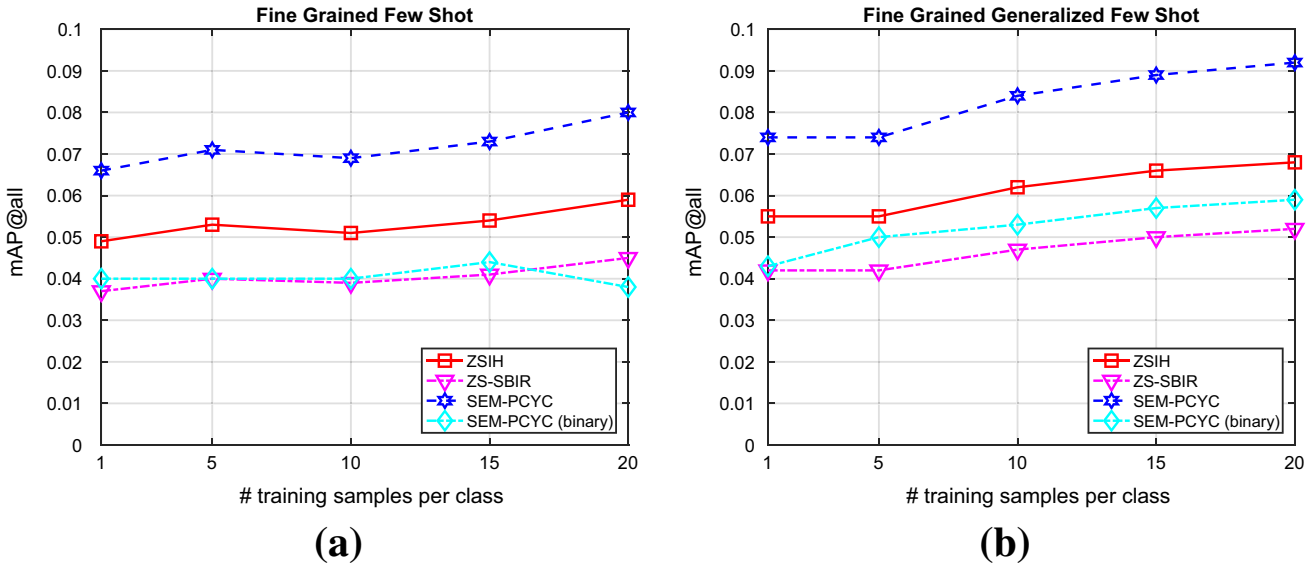

Fig. 12 Fine-grained (generalized) few-shot sketch-based image retrieval performance comparison (color figure online)

sketch and image to the same neighborhood and ignores fine-grained information. Therefore the proposed solution to low-shot task and the notion of fine-grained problem contradicts, and as a consequence the performance of all the considered models deteriorates. In generalized setting, we have observed that all the models have performed slightly better. We conjecture that the considered models can memorize the fine-grained information of the training or *seen* samples, which gives a slight rise (as they are very few in number) in performance in generalized scenario. However, we see that low-shot fine-grained paradigm is very important for SBIR. Nevertheless, we admit that it is an extremely challenging task, which needs substantial research work to be solved.

5 Conclusion

In this paper, we proposed the SEM-PCYC model for the any-shot SBIR task. Our SEM-PCYC model is a semantically aligned paired cycle consistent generative adversarial network 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. In addition to the model, in this paper, we introduced (generalized) few-shot SBIR as a new task, which is combined with fine-grained setting. We considered three benchmark datasets with varying difficulties and challenges, and performed exhaustive evaluation with the above mentioned paradigms. Our assessment on these three datasets has shown that our model consistently outperforms the existing methods in (generalized) zero- and few-shot, and fine-grained settings. We encourage future work to evaluate sketch based image retrieval methods in these incrementally challenging and realistic settings.

Acknowledgements This work has received funding from the European Union under Marie Skłodowska-Curie Grant Agreement No. 665919, from the ERC under the Horizon 2020 program (Grant Agreement No. 853489), the Spanish Ministry project RTI2018-102285-A-I00 and DFG-EXC-Nummer 2064/1-Projektnummer 390727645. The TITAN Xp and TITAN V used for this research were donated by the NVIDIA Corporation.

References

Akata, Z., Malinowski, M., Fritz, M., & Schiele, B. (2016). Multi-cue zero-shot learning with strong supervision. In *CVPR* (pp. 59–68).

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

Akata, Z., Reed, S., Walter, D., Lee, H., & Schiele, B. (2015). Evaluation of output embeddings for fine-grained image classification. In *CVPR* (pp. 2927–2936).

Al-Halah, Z., Tapaswi, M., & Stiefelhagen, R. (2016). Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR* (pp. 5975–5984).

Changpinyo, S., Chao, W., Gong, B., & Sha, F. (2016). Synthesized classifiers for zero-shot learning. In *CVPR* (pp. 5327–5336).

Changpinyo, S., Chao, W., & Sha, F. (2017). Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV* (pp. 3496–3505).

Chen, J., & Fang, Y. (2018). Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In *ECCV* (pp. 624–640).

Chen, L., Zhang, H., Xiao, J., Liu, W., & Chang, S. (2018). Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR* (pp. 1043–1052).

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *CVPR* (pp. 539–546).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR* (pp. 248–255).

Dey, S., Riba, P., Dutta, A., Lladós, J., & Song, Y. Z. (2019). Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*.

Ding, Z., Shao, M., & Fu, Y. (2017). Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR* (pp. 6005–6013).

Dutta, A., & Akata, Z. (2019). Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*. (pp. 5084–5093)

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM TG*, *31*(4), 1–10.

Felix, R., Kumar, V. B. G., Reid, I., & Carneiro, G. (2018). Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV* (pp. 21–37).

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML* (pp. 1126–1135).

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., et al. (2013). Devise: A deep visual-semantic embedding model. In *NIPS* (pp. 2121–2129).

Fu, Z., Xiang, T., Kodirov, E., & Gong, S. (2015). Zero-shot object recognition by semantic manifold distance. In *CVPR* (pp. 2635–2644).

Girshick, R. (2015). Fast r-cnn. In *ICCV* (pp. 1440–1448).

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

Guo, Y., Ding, G., Han, J., & Tang, S. (2018). Zero-shot learning with attribute selection. In *AAAI* (pp. 6870–6877).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv:1512.03385

Hu, G., Hua, Y., Yuan, Y., Zhang, Z., Lu, Z., Mukherjee, S. S., et al. (2017). Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV* (pp. 3764–3773).

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-excitation networks. In *IEEE TPAMI* (pp. 2011–2023).

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

Jayaraman, D., & Grauman, K. (2014). Zero-shot recognition with unreliable attributes. In *NIPS* (pp. 3464–3472).

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING* (pp. 19–33).

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2017). Fasttext.zip: Compressing text classification models. In *ICLR* (pp. 1–13).

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR* (pp. 1–10).

Kiran Yelamarthi, S., Krishna Reddy, S., Mishra, A., & Mittal, A. (2018). A zero-shot framework for sketch based image retrieval. In *ECCV* (pp. 316–333).

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML DLW* (pp. 1–8).

Kodirov, E., Xiang, T., & Gong, S. (2017). Semantic autoencoder for zero-shot learning. In *CVPR* (pp. 4447–4456).

Lampert, C. H., Nickisch, H., & Harmeling, S. (2014). Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, *36*(3), 453–465.

Li, Y., Hospedales, T. M., Song, Y. Z., & Gong, S. (2014). Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC* (pp. 1–12).

Lin, D. (1998). An information-theoretic definition of similarity. In *ICML* (pp. 296–304).

Liu, L., Shen, F., Shen, Y., Liu, X., & Shao, L. (2017). Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR* (pp. 2298–2307).

Liu, Q., Xie, L., Wang, H., & Yuille, A. L. (2019). Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV* (pp. 3661–3670).

Long, Y., Liu, L., Shao, L., Shen, F., Ding, G., & Han, J. (2017). From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *CVPR* (pp. 6165–6174).

Mensink, T., Gavves, E., & Snoek, C. G. M. (2014). Costa: Co-occurrence statistics for zero-shot classification. In *CVPR* (pp. 2441–2448).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR* (pp. 1–12).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS* (pp. 3111–3119).

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

Nie, F., Huang, H., Cai, X., & Ding, C. H. (2010). Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS* (pp. 1813–1821).

Pang, K., Li, K., Yang, Y., Zhang, H., Hospedales, T. M., Xiang, T., et al. (2019). Generalising fine-grained sketch-based image retrieval. In *CVPR* (pp. 677–686).

Pang, K., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2017). Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC* (pp. 1–12).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in PyTorch. In *NIPS-W* (pp. 1–12).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP* (pp. 1532–1543).

Qi, Y., Song, Y. Z., Zhang, H., & Liu, J. (2016). Sketch-based image retrieval via siamese convolutional neural network. In *ICIP* (pp. 2460–2464).

Qiao, R., Liu, L., Shen, C., & Van Den Hengel, A. (2016). Less is more: Zero-shot learning from online textual documents with noise suppression. In *CVPR* (pp. 2249–2257).

Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. In *ICLR* (pp. 1–12).
Reed, S., Akata, Z., Lee, H., & Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. In *CVPR* (pp. 49–58).
Romera-Paredes, B., & Torr, P. H. S. (2015). An embarrassingly simple approach to zero-shot learning. In *ICML* (pp. 2152–2161).
Saavedra, J. M. (2014). Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *ICIP* (pp. 2998–3002).
Saavedra, J. M., & Barrios, J. M. (2015). Sketch based image retrieval using learned keyshapes (lks). In *BMVC* (pp. 1–11).
Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. *ACM TOG*, *35*(4), 1–12.
Satorras, V. G., & Estrach, J. B. (2018). Few-shot learning with graph neural networks. In *ICLR* (pp. 1–13).
Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., & Akata, Z. (2018) Generalized zero- and few-shot learning via aligned variational autoencoders. In *CVPR* (pp. 8247–8255).
Shen, Y., Liu, L., Shen, F., & Shao, L. (2018). Zero-shot sketch-image hashing. In *CVPR* (pp. 3598–3607).
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.*, *30*, 4077–4087.
Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *NIPS* (pp. 935–943).
Song, J., Song, Y. Z., Xiang, T., & Hospedales, T. (2017a). Fine-grained image retrieval: The text/sketch input dilemma. In *BMVC* (pp. 1–12).
Song, J., Yu, Q., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2017b). Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV* (pp. 5552–5561).
Su, W., Yuan, Y., & Zhu, M. (2015). A relationship between the average precision and the area under the roc curve. In *ICTIR* (pp. 349–352).
Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., & Wierstra, D. (2016). Matching networks for one shot learning. In *NIPS* (pp. 3630–3638).
Wang, F., Kang, L., & Li, Y. (2015a). Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR* (pp. 1875–1883).
Wang, M., Wang, C., Yu, J. X., & Zhang, J. (2015b). Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework. In *VLDB* (pp. 998–1009).
Wang, S., Ding, Z., & Fu, Y. (2017). Feature selection guided auto-encoder. In *AAAI* (pp. 2725–2731).
Wang, W., Pu, Y., Verma, V. K., Fan, K., Zhang, Y., Chen, C., Rai, P., & Carin, L. (2018a). Zero-shot learning via class-conditioned deep generative models. In *AAAI* (pp. 4211–4218).
Wang, Y., Girshick, R., Hebert, M., & Hariharan, B. (2018b). Low-shot learning from imaginary data. In *CVPR* (pp. 7278–7286).
Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., & Schiele, B. (2016). Latent embeddings for zero-shot classification. In *CVPR* (pp. 69–77).
Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018a). Zero-shot learning–a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, *41*(9), 2251–2265.
Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. (2018b). Feature generating networks for zero-shot learning. In *CVPR* (pp. 5542–5551).
Xian, Y., Sharma, S., Schiele, B., & Akata, Z. (2019). f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR* (pp. 10275–10284).
Yang, Y., Luo, Y., Chen, W., Shen, F., Shao, J., & Shen, H. T. (2016a). Zero-shot hashing via transferring supervised knowledge. In *ACM MM* (pp. 1286–1295).
Yang, Z., Cohen, W. W., & Salakhutdinov, R. (2016b). Revisiting semi-supervised learning with graph embeddings. In *ICML* (pp. 40–48).
Yu, Q., Liu, F., Song, Y. Z., Xiang, T., Hospedales, T. M., & Loy, C. C. (2016a). Sketch me that shoe. In *CVPR* (pp. 799–807).
Yu, Q., Yang, Y., Liu, F., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2016b). Sketch-a-net: A deep neural network that beats humans. *IJCV*, 122, 411–425
Yu, Q., Yang, Y., Song, Y. Z., Xiang, T., & Hospedales, T. (2015). Sketch-a-net that beats humans. In *BMVC*, pp. 1–12.
Yu, T., Meng, J., & Yuan, J. (2018). Multi-view harmonized bilinear network for 3d object recognition. In *CVPR* (pp. 186–194).
Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV* (pp. 1839–1848).
Zhang, J., Shen, F., Liu, L., Zhu, F., Yu, M., Shao, L., et al. (2018). Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV* (pp. 304–321).
Zhang, L., Xiang, T., & Gong, S. (2017). Learning a deep embedding model for zero-shot learning. In *CVPR* (pp. 3010–3019).
Zhang, R., Lin, L., Zhang, R., Zuo, W., & Zhang, L. (2015). Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE TIP*, *24*(12), 4766–4779.
Zhang, Z., & Saligrama, V. (2015). Zero-shot learning via semantic similarity embedding. In *ICCV* (pp. 4166–4174).
Zhang, Z., & Saligrama, V. (2016). Zero-shot learning via joint latent similarity embedding. In *CVPR* (pp. 6034–6042).
Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV* (pp. 2242–2251).