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 have 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 auto-encoder 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.

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 aligns the sketch and image in the semantic space. In- fact, we train a mapping of 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 \( G_{sk} : X \rightarrow S \) so that \( \hat{s}_i = G_{sk}(x_i) \), where \( s_i \in S \) is the corresponding side information and is indistinguishable from \( s_i \) via an adversarial training that classifies \( \hat{s}_i \) different from \( s_i \). The optimal \( G_{sk} \) thereby translates the modality \( X \) into a modality \( S \) which is identically distributed to \( S \). Similarly, another function \( G_{im} : Y \rightarrow S \) can be trained via the same discriminator such that \( \hat{s}_i = G_{im}(y_i) \).

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

\[
L_{adv}(G_{sk}, G_{im}, D_{se}, x, y, s) = 2 \times \mathbb{E} \left[ \log D_{se}(s) \right] - \mathbb{E} \left[ \log(1 - D_{se}(G_{sk}(x))) \right] - \mathbb{E} \left[ \log(1 - D_{se}(G_{im}(y))) \right]
\]

where \( G_{sk} \) and \( G_{im} \) generate side information similar to the ones in \( S \) 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., \( \min_{G_{sk}, G_{im}} \max_{D_{se}} L_{adv}(G_{sk}, G_{im}, D_{se}, x, y, s) \). In a similar way, for the generator \( F_{sk} \) and its discriminator \( D_{sk} \), the objective is:

\[
L_{adv}(F_{sk}, D_{sk}, x, s) = \mathbb{E} \left[ \log D_{sk}(x) \right] + \mathbb{E} \left[ \log(1 - D_{sk}(F_{sk}(s))) \right]
\]
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. \( \min F_{sk} \max D_{sk} \mathcal{L}_{adv}(F_{sk}, D_{sk}, x, s) \). Similarly, another adversarial loss is introduced for the mapping \( F_{im} \) and its discriminator \( D_{im} \), i.e., \( \min F_{im} \max D_{im} \mathcal{L}_{adv}(F_{im}, D_{im}, y, s) \).

**Cycle Consistency Loss.** The adversarial mechanism effectively reduces the domain or modality gap, however, it is not guaranteed that an input \( x_i \) and an output \( s_i \) 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 \( G_{sk} : X \rightarrow S \) and another mapping \( F_{sk} : S \rightarrow X \), then both \( G_{sk} \) and \( F_{sk} \) are reverse of each other, and hence form a one-to-one correspondence or bijective mapping.

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

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

**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.

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

where \( c \) is the category label of \( x \). 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 \( s \in \mathbb{R}^k \) be the side information with \( k \) as the original dimension. The loss function is:

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

where \( f(s) = \sigma(W_1s + b_1) \), \( g(f(s)) = \sigma(W_2f(s) + b_2) \), 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:

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

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-PYC 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 code important semantic information. The objective is to code important semantic information. The objective is to

\[L_{adv}(G_{sk}, G_{im}, D_{se}, x, y, s) = \mathcal{L}_{adv}(G_{sk}) + \mathcal{L}_{adv}(G_{im}) + \mathcal{L}_{aenc}(f, g)\]

1. https://wordnet.princeton.edu

2. Our code and trained models are available at: https://github.com/AnjanDutta/sem-pyc

5093
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], DeiSE [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-

![Swan, duck, owl, penguin, standing bird](image)

Figure 3. Inter-class similarity in TU-Berlin dataset.
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)

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

<table>
<thead>
<tr>
<th>Text Embedding</th>
<th>Hierarchical Embedding</th>
<th>Sketchy (Extended)</th>
<th>TU-Berlin (Extended)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.284</td>
<td>0.228</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.330</td>
<td>0.232</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.314</td>
<td>0.224</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.246</td>
<td>0.169</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.308</td>
<td>0.227</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.338</td>
<td>0.276</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.299</td>
<td>0.253</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.285</td>
<td>0.243</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.340</td>
<td>0.297</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.288</td>
<td>0.264</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>0.349</td>
<td>0.291</td>
</tr>
</tbody>
</table>

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
modalities might be matched in wrong way. Hence, it can be learned embedding could be very discriminative but the two retrieval results. We conjecture that in this case the consideration criteria together with adversarial loss, neither improves the performance of our system does not improve substantially 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 $\ell_{21}$-norm based regularizer that effectively select representative side information.

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

<table>
<thead>
<tr>
<th>Description</th>
<th>Sketchy</th>
<th>TU-Berlin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only adversarial loss</td>
<td>0.128</td>
<td>0.109</td>
</tr>
<tr>
<td>Adversarial + cycle consistency loss</td>
<td>0.147</td>
<td>0.131</td>
</tr>
<tr>
<td>Adversarial + classification loss</td>
<td>0.140</td>
<td>0.127</td>
</tr>
<tr>
<td>Without selecting side information</td>
<td>0.382</td>
<td>0.299</td>
</tr>
<tr>
<td>Without regularizer in eqn. (5)</td>
<td>0.323</td>
<td>0.273</td>
</tr>
<tr>
<td><strong>SEM-PCYC (full model)</strong></td>
<td><strong>0.349</strong></td>
<td><strong>0.297</strong></td>
</tr>
</tbody>
</table>

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.

### 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

This work has been partially supported by European Union’s research and innovation program under Marie Skłodowska-Curie grant agreement No. 665919. The Titan Xp and Titan V used for this research were donated by the NVIDIA Corporation.
References

[33] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In BMVC, pages 1–12, 2017. 1, 2