A Zero-Shot Framework for Sketch Based Image Retrieval.

Sasi Kiran Yelamarthi*, Shiva Krishna Reddy*, Ashish Mishra, and Anurag Mittal

Indian Institute of Technology Madras, India
{sasikiran1996, shivakrishnam912}@gmail.com, {mishra, amittal}@cse.iitm.ac.in

Abstract. Sketch-based image retrieval (SBIR) is the task of retrieving images from a natural image database that correspond to a given hand-drawn sketch. Ideally, an SBIR model should learn to associate components in the sketch (say, feet, tail, etc.) with the corresponding components in the image having similar shape characteristics. However, current evaluation methods simply focus only on coarse-grained evaluation where the focus is on retrieving images which belong to the same class as the sketch but not necessarily having the same shape characteristics as in the sketch. As a result, existing methods simply learn to associate sketches with classes seen during training and hence fail to generalize to unseen classes. In this paper, we propose a new benchmark for zero-shot SBIR where the model is evaluated on novel classes that are not seen during training. We show through extensive experiments that existing models for SBIR that are trained in a discriminative setting learn only class specific mappings and fail to generalize to the proposed zero-shot setting. To circumvent this, we propose a generative approach for the SBIR task by proposing deep conditional generative models that take the sketch as an input and fill the missing information stochastically. Experiments on this new benchmark created from the "Sketchy" dataset, which is a large-scale database of sketch-photo pairs demonstrate that the performance of these generative models is significantly better than several state-of-the-art approaches in the proposed zero-shot framework of the coarse-grained SBIR task.

Keywords: Image Retrieval, Zero-Shot Learning

1 Introduction

The rise in the number of internet users coupled with increased storage capacity, better internet connectivity and higher bandwidths has resulted in an exponential growth in multimedia content on the Web. In particular, image content has become ubiquitous and plays an important role in engaging users on social media as well as customers on various e-commerce sites. With this growth in image content, the information needs and search patterns of users have also evolved. Specifically, it is now common for users to search for images (instead of documents) either by providing a textual description of the image or by providing

*: Equal Contribution

another image which is similar to the desired image. The former is known as text based image retrieval and the latter as content based image retrieval [18].

The motivation for content based image retrieval can be easily understood by taking an example from online fashion. Here, it is often hard to provide a *textual description* of the desired product but easier to provide a *visual description* in the form of a matching image. The visual description/query need not necessarily be an image but can also be a sketch of the desired product, if no image is available. The user can simply draw the sketch on-the-fly on touch based devices. This convenience in expressing a visual query has led to the emergence of Sketch-based image retrieval (SBIR) as an active area of research [3–5, 9, 13, 15, 17, 24, 30, 31, 36, 38, 47, 51, 54]. The primary challenge here is the domain gap between images and sketches wherein sketches contain only an outline of the object and hence have less information compared to images. The second challenge is the large intra-class variance present in sketches due to the fact that humans tend to draw sketches with varied levels of abstraction. Ideally, for better generalization, a model for SBIR must learn to discover the alignments between the components of the sketch and the corresponding image. For example, in Figure-1, we would want the model to associate the head of the cow in the sketch to that in the image. However, current evaluation methodology [7, 26, 40] focuses only on class-based retrieval rather than shape or attribute-based retrieval. Specifically, during evaluation, the model is given credit if it simply fetches an image which belongs to the same class as the sketch. The object in the image need not have the same outline, etc as in the sketch. For example, for the query (sketch) shown in Figure-1, there is no guarantee that the model fetches the image of the cow with the same number of feet visible or the tail visible, even if it has high evaluation score.

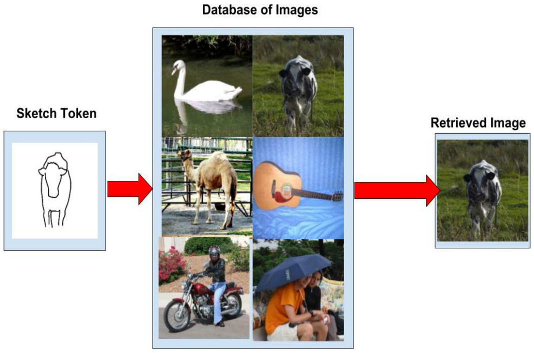

Fig. 1. Illustration of Sketch based Image Retrieval

Thus, a model could possibly achieve good performance by simply learning a class specific mapping from sketches to class labels and retrieve all the images from the same class as that of the query sketch. This is especially so, when the unseen sketches seen at test time belong to the same set of classes as seen during training. Furthermore, existing methods evaluate their models on a set of randomly selected sketches that are withheld during training. However, the images corresponding to the withheld sketches could still occur in the training set, and that would make the task easier.

One way to discourage such class specific learning is to employ a fine-grained evaluation [30, 51]. For a given sketch, the retrieved results are evaluated by comparing the estimated ranking of images in the database with a human annotated rank list. However, creating such annotations for large datasets such as "Sketchy" [40] requires extensive human labor. Also, such evaluation metrics are subject to human biases. In this work, we propose coarse-grained evaluation in the zero-shot setting as a surrogate to fine-grained evaluation to circumvent both these drawbacks. The idea is to test the retrieval on sketches of unseen classes to discourage class-specific learning during training. The evaluation is automatic, *i.e.*, it requires no human labor for each retrieval, apart from having no biases. The model has to learn to associate the latent alignments in the sketch and the image in order to perform well. This is also important from a practical standpoint wherein, in some domains, all possible classes many not be available at training time. For example, new product classes emerge every day in the fashion industry. Thus, the *Zero-Shot Sketch Based Image Retrieval (ZS-SBIR)* task introduced in this paper provides a more realistic setup for the sketch-based retrieval task.

Towards this end, we propose a new benchmark for the ZS-SBIR task by creating a careful split of the Sketchy database. We first evaluate several existing SBIR models on this task and observe that the performance of these models drops significantly in the zero-shot setting thereby pointing to class-specific learning occurring in these models. We hypothesize that one reason for this could be that the existing methods are essentially formulated in the discriminative setup, which encourages class specific learning. To circumvent the problems in these existing models, we approach the problem from the point of view of a generative model. Specifically, ZS-SBIR can be considered as the task of generating additional information that is absent in the sketch in order to retrieve similar images. We propose Deep Conditional Generative Models based on Adversarial Autoencoders and Variational Autoencoders for the ZS-SBIR task. Our experiments show that the proposed generative approach performs better than all existing state-of-the-art SBIR models in the zero-shot setting.

The paper is organized as follows: In Section-2, we give a brief overview of the state-of-the-art techniques in SBIR and ZSL. Subsequently, in Section-3, we introduce the proposed zero-shot framework and describe the proposed dataset split. Section-4 shows the evaluation of existing state-of-the-art SBIR models in this proposed setting. Section-5 introduces our proposed generative modeling of ZS-SBIR and adaptations of three popular ZSL models to this setting. Finally, in Sections-6, we present an empirical evaluation of these models on the proposed zero shot splits on the Sketchy dataset.

2 Related Work

Since we propose a zero-shot framework for the SBIR task, we briefly review the literature from both sketch-based image retrieval as well as zero-shot learning in this section.

Conventional pipeline in SBIR involves projecting images and sketches into a common feature space. These features or binary codes extracted from them are used for the retrieval task. Hand-crafted feature based models include the gradient field HOG descriptor proposed by Hu and Collomose [14], the histogram of edge orientations (HELO) proposed by Saavendra [37], the learned key shapes (LKS) proposed by Saavendra *et.al* [39] which are used in Bag of Visual Words (BoVW) framework as feature extractors for SBIR. Yu *et.al* [52] were the first to use Convolutional Neural Networks (CNN) for the sketch classification task. Qi *et.al* [7] introduced the use of siamese architecture for coarse-grained SBIR. Sangkloy *et.al* [40] used triplet ranking loss for training the features for coarse-grained SBIR. Yu *et.al* [51] used triplet network for instance level SBIR evaluating the performance on shoe and chair dataset. They use a pseudo fine-grained evaluation where they only look at the position of the correct image for a sketch in the retrieved images. Liu *et.al* [26] propose a semi-heterogeneous deep architecture for extracting binary codes from sketches and images that can be trained in an end-to-end fashion for coarse-grained SBIR task.

We now review the zero-shot literature. Zero-shot learning in Image Classification [22, 23, 28] refers to learning to recognize images of novel classes although no examples from these classes are present in the training set. Due to the difficulty in collecting examples of every class in order to train supervised models, zero-shot learning has received significant interest from the research community recently [1, 10, 21, 23, 35, 43, 46, 48, 49]. We refer the reader to [50] for a comprehensive survey on the subject. Recently, zero shot learning has been gaining increasing attention for a number of other computer vision tasks such as image tagging [25, 53], visual question answering [29, 33, 45] etc. To the best of our knowledge, the zero-shot framework has not been previously explored in the SBIR task.

3 Zero shot setting for SBIR

We now provide a formal definition of the zero shot setting in SBIR. Let $S = \{(x_i^{sketch}, x_i^{img}, y_i) | y_i \in \mathcal{Y}\}$ be the triplets of sketch, image and class label where $\mathcal{Y}$ is the set of all class labels in S. We partition the class labels in the data into $\mathcal{Y}_{train}$ and $\mathcal{Y}_{test}$ data respectively. Correspondingly, let $S_{tr} = \{(x_i^{sketch}, x_i^{img}) | y_i \in Y_{train}\}$ and $S_{te} = \{(x_i^{sketch}, x_i^{img}) | y_i \in Y_{test}\}$ be the partition of S into train and test sets. This way, we partition the paired data into train and test set such that none of the sketches from the test classes occur in the train set. Since the model has no access to class labels, the model needs to learn latent alignments between the sketches and the corresponding images to perform well on the test data.

Let D be the database of all images and g_I be the mapping from images to class labels. We split D into $D_{tr} = \{x_i^{img} \in D | g_I(x_i^{img}) \in Y_{train}\}$ and $D_{te} = \{x_i^{img} \in D | g_I(x_i^{img}) \in Y_{test}\}$. This is similar to other zero-shot literature [23] in image classification. The retrieval model in this framework can only be trained on S_{tr}. The database D_{tr} may be used for validating the retrieval results in order to tune the hyper-parameters. Given an x^{sketch} taken from sketches of S_{te}, the objective of zero shot setting in SBIR is to retrieve images from D_{te} that belong to same class as that of the query sketch. This evaluation setting ensures that the model can not just learn the mapping from sketches to class labels and retrieve all the images using the label information. The model now has to learn the salient common features between sketches and images and use this to retrieve images for the query that are from the unseen classes.

3.1 Benchmark

Since we are introducing the task of zero-shot sketch based retrieval, there is no existing benchmark for evaluating this setting. Hence, we first propose a new benchmark for evaluation by making a careful split of the ”Sketchy” dataset [40]. Sketchy is a dataset consisting of 75,471 hand-drawn sketches and 12,500 images belonging to 125 classes collected by Sangkloy *et.al* [40]. Each image has approximately 6 hand-drawn sketches. The original Sketchy dataset uses the same 12,500 images as the database. Liu *et.al* [26] augment the database with 60,502 images from Imagenet to create a retrieval database with a total of 73,002 images. We use the augmented dataset provided by Liu *et.al* [26] in this work.

Next, we partition the 125 classes into 104 train classes and 21 test classes. This peculiar split is not arbitrary. We make sure that the 21 test classes are not present in the 1000 classes of Imagenet [8]. This is done to ensure that researchers can still pre-train their models on the 1000 classes of Imagenet without violating the zero-shot assumption. Such a split was motivated by the recently proposed benchmark for standard datasets used in the zero shot image classification task by Xian *et.al* [50]. The details of the proposed dataset split are summarized in Table 1.

Table 1. Statistics of the proposed dataset split of Sketchy database for ZS-SBIR task

Dataset Statistics	#
Train classes	104
Test classes	21
Train Images	10400
Train Sketches	62787
Avg. sketches per image	6.03848
Test Sketches	12694
DB images for training	62549
DB images for testing	10453

4 Limitations of existing SBIR methods

Next we evaluate whether the existing approaches to the sketch-based image retrival task generalize well to the proposed zero-shot setting. To this end, we evaluate three state-of-the-art SBIR methods described below on the above proposed benchmark.

4.1 A Siamese Network

The Siamese network proposed by Hadsell *et.al* [12] maps both the sketches and images into a common space where the semantic distance is preserved. Let $(S, I, Y = 1)$ and $(S, I, Y = 0)$ be the pairs of images and sketches that belong to same and different class respectively and $D_\theta(S, I)$ be the l2 distance between the image and sketch features where θ are the parameters of the mapping function. The loss function $L(\theta)$ for training is given by:

$$L(\theta) = (Y)\frac{1}{2}(D_\theta)^2 + (1-Y)\frac{1}{2}\{max(0, m - D_\theta)\}^2 \tag{1}$$

where m is the margin. Chopra *et.al* [7] and Qi *et.al* [32] use a modified version of the above loss function for training the Siamese network for the tasks of face verification and SBIR respectively, which is given below:

$$L(\theta) = (Y)\alpha D_\theta^2 + (1-Y)\beta e^{\gamma D_\theta} \tag{2}$$

where $\alpha = \frac{2}{Q}$, $\beta = 2Q$, $\gamma = -\frac{2.77}{Q}$ and constant Q is set to the upper bound on D_θ estimated from the data. We explore both these formulations in the proposed zero-shot setting. We call the former setting as Siamese-1 and the latter as Siamese-2.

4.2 A Triplet Network

Triplet loss [40, 41] is defined in a max-margin framework, where, the objective is to minimize the distance between sketch and positive image that belong to the same class and simultaneously maximize the distance between the sketch and negative image which belong to different classes. The triplet training loss for a given triplet $t(s, p^+, p^-)$ is given by:

$$L_\theta(t) = max(0, m + D_\theta(s, p^+) - D_\theta(s, p^-)) \tag{3}$$

where m is the margin and D_θ is the distance measure used.

To sample the negative images during training, we follow two strategies (i) we consider only images from different class and (ii) we consider all the images that do not directly correspond to the sketch, resulting in coarse-grained and fine-grained training of triplet network respectively. We explore both these training methods in the proposed zero-shot setting for SBIR.

Table 2. Precision and mAP are estimated by retrieving 200 images. - indicates that the authors do not present results on that metric. 1:Using 128 bit hash codes

Method	**Precision@200**		**mAP@200**	
	Traditional	**Zero-Shot**	**Traditional**	**Zero-Shot**
Baseline	-	0.106	-	0.054
Siamese-1	-	0.243	-	0.134
Siamese-2	0.690	0.251	0.518	0.149
Coarse-grained triplet	0.761	0.169	0.573	0.083
Fine-grained triplet	-	0.155	-	0.081
DSH[1]	0.866	0.153	0.783	0.059

4.3 Deep Sketch Hashing(DSH)

Liu *et.al* [26] propose an end-to-end framework for learning binary codes of sketches and images which is the current state-of-the-art in SBIR. The objective function consists of the following three terms: (i) cross-view pairwise loss which tries to bring binary codes of images and sketches of the same class to be close (ii) semantic factorization loss which tries to preserve the semantic relationship between classes in the binary codes and (iii) the quantization loss.

4.4 Experiments

We now present the results of the above described models on our proposed partitions of the ”Sketchy” dataset [40] in order to evaluate them in the zero-shot setting.

While evaluating each model, for a given test sketch, we retrieve the top $K = 200$ images from the database that are closest to the sketch in the learned feature space. We use inverse of the cosine similarity as the distance metric. We present the experimental details for the evaluated methods below.

Baseline: We take a VGG-16 network [42] trained on image classification task on ImageNet-1K [8] as the baseline. The score for a given sketch-image pair is given by the cosine similarity between their VGG features.

Training: We re-implement the above described models to evaluate them for the ZS-SBIR task. For sanity check, we first reproduce the results on the traditional SBIR task reported in [26] successfully. We follow the training methodology described in [7, 26, 40] closely.

We observe that the validation performance saturates after 20 epochs in case of Siamese network and after 80 epochs for the Triplet network. We also employ data augmentation for training the Triplet network because the available training data is insufficient for proper training. We explore the hyper-parameters via grid search.

In the case of DSH, we use the CNNs proposed by Liu *et.al* [26] for feature extraction. We train the network for 500 epochs, validating on the train database after every 10 epochs. We explored the hyper-parameters and found that $\lambda = 0.01$ and $\gamma = 10^{-5}$ give the best results similar to the original SBIR training.

The performance of these models on the ZS-SBIR task are shown in Table 2. For comparative purposes, we also present the performance in the traditional SBIR setting [26] where the models are trained on the sketch-image pairs of all the classes. We observe that the performance of these models dips significantly, indicating the non-generalizability of existing approaches to SBIR. This performance drop of more than 50% in the zero-shot setting may be due to the fact that these models trained in a discriminative setting may learn to associate the sketches and images to class labels.

Among the compared methods we notice that the Siamese network preforms the best among the existing SBIR methods in the zero-shot setting. We also observe that the Triplet loss gives poorer performance compared to the Siamese network. This can be attributed to the presence of only about $60,000$ images during training, which is not sufficient for properly training a triplet network as observed by Schroff *et.al* [41]. We also observe that the coarse-grained training of triplet performs better compared to fine-grained triplet. This may be because the fine-grained training considers all the images other than those that correspond directly to the sketch as negative samples making the training harder.

Our next observation is that DSH, which is the state-of-the-art model in SBIR does not perform well compared to either Siamese or Triplet networks in ZS-SBIR task. This may be due to the fact that the semantic factorization loss in DSH takes only the training class embeddings into account and does not reduce the semantic gap for the test classes.

Thus, one can claim that there exists a problem of class-based learning inherent in the existing models, which leads to inferior performance in the ZS-SBIR task.

5 Generative Models for ZS-SBIR

Having noticed that the existing approaches do not generalize well to the ZS-SBIR task, we now propose the use of generative models for the ZS-SBIR task. The motivation for such an approach is that while a sketch gives a basic outline of the image, additional details could possibly be generated from the latent prior vector via a generative model. This is inline with the recent work on similar image translation tasks [6, 16, 34] in computer vision.

Let G_θ model the probability distribution of the image features (x_{img}) conditioned on the sketch features (x_{sketch}) and parameterized by θ, i.e $\mathbb{P}(x_{img}|x_{sketch};\theta)$. G_θ is trained using paired data of sketch-image pairs from the training classes. Since we do not provide the model with class label information, it is hoped that the model learns to associate the characteristics of the sketch such as the general outline, local shape, etc with that of the image. We would like to emphasize here that G_θ is trained to generate image features but not the images themselves using the sketch. We consider two popular generative models: Variational Autoencoders [20, 44] and Adversarial Autoencoders [27] as described below:

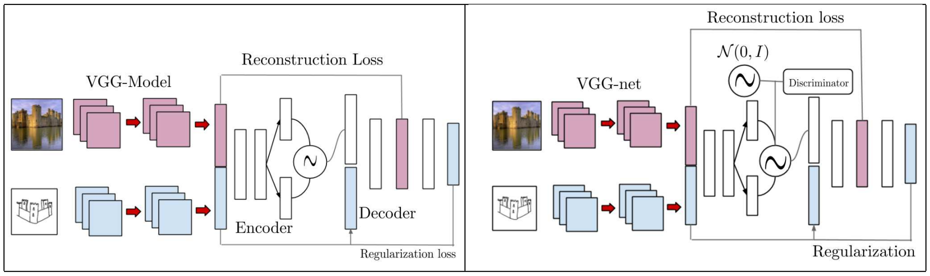

Fig. 2. The architectures of CVAE and CAAE are illustrated in the left and right diagrams respectively

5.1 Variational Autoencoders

The Variational Autoencoders (VAE) [20] map a prior distribution on a hidden latent variable $p(z)$ to the data distribution $p(x)$. The intractable posterior $p(z|x)$ is approximated by the variational distribution $q(z|x)$ which is assumed to be Gaussian in this work. The parameters of the variational distribution are estimated from x via the encoder which is a neural network parameterized by ϕ. The conditional distribution $p(x|z)$ is modeled by the decoder network parameterized by θ. Following the notation in [20], the variational lower bound for $p(x)$ can be written as:

$$\begin{aligned} p(x) &\geq \mathcal{L}(\phi, \theta; x) \\ &= -D_{KL}\left(q_\phi(z|x)||p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \end{aligned} \tag{4}$$

Similarly, it is possible to model the conditional probability $p(x|y)$ as proposed by [44]. In this work, we model the probability distribution over images conditioned on the sketch i.e $P\left(x_{img}|x_{sketch}\right)$. The bound now becomes:

$$\mathcal{L}(\phi, \theta; x_{img}, x_{sketch}) =$$

$$\begin{aligned} -D_{KL}\left(q_\phi\left(z|x_{img}, x_{sketch}\right)||p_\theta\left(z|x_{sketch}\right)\right) + \\ \mathbb{E}\left[\log p_\theta\left(x_{img}|z, x_{sketch}\right)\right] \end{aligned} \tag{5}$$

Furthermore, to encourage the model to preserve the latent alignments of the sketch, we add the reconstruction regularization to the objective. In other words, we force the reconstructibility of the sketch features from the generated image features via a one-layer neural network f_{NN} with parameters ψ. All the parameters $\theta, \psi \& \phi$ are trained end-to-end. The regularization loss can be expressed as

$$\mathcal{L}_{recons} = \lambda . \, ||f_{NN}(\widehat{x}_{img}) - x_{sketch}||_2^2 \tag{6}$$

Here, λ is a hyper-parameter which is to be tuned. The architecture of the conditional variational autoencoder used is shown in Figure 2. We call this CVAE from here on.

5.2 Adversarial Autoencoders

Adversarial Autoencoders [27] are similar to the variational autoencoder, where the KL-Divergence term is replaced with an adversarial training procedure. Let E, D be the encoder and decoder of the autoencoder respectively. E maps input x_{img} to the parameters of the hidden latent vector distribution $P(z|x_{img})$, whereas, D maps the sampled z to x_{img} (both are conditioned on the sketch vector x_{sketch}). We have an additional network $\mathcal{D}$: the discriminator. The networks E & D try to minimize the following loss:

$$\mathbb{E}_z\left[\log p_\theta\left(x_{img}|z, x_{sketch}\right)\right] + \mathbb{E}_{x_{img}}\left[\log\left(1 - \mathcal{D}(E(x_{img}))\right)\right] \tag{7}$$

The discriminator $\mathcal{D}$ tries to maximize the following similar to the original GAN formulation [11]:

$$\mathbb{E}_z\left[\log\left[\mathcal{D}(z)\right]\right] + \mathbb{E}_{x_{img}}\left[\log\left[1 - \mathcal{D}\left(E(x_{img})\right)\right]\right] \tag{8}$$

We add the reconstructibility regularization described in the above section to the loss of the encoder. The architecture of the adversarial autoencoder used is shown in Figure 2. We call this CAAE from here on.

5.3 Retrieval Methodology

G_θ is trained on the sketch-image feature pairs from the seen classes. During test time, the decoder part of the network is used to generate a number of image feature vectors x^I_{gen} conditioned on the test sketch by sampling latent vectors from the prior distribution $p(z) = \mathcal{N}(0, I)$. For a test sketch x_S corresponding to a test class, we generate the set $\mathcal{I}_{x_S}$ consisting of N (a hyper-parameter) such samples of x^I_{gen}. We then cluster these generated samples $\mathcal{I}_{x_S}$ using K-Means clustering and obtain K cluster centers $C_1, C_2, \ldots, C_k$ for each test sketch. We retrieve 200 images x^I_{db} from the image database based on the following distance metric:

$$\mathcal{D}(x^{db}_I, \mathcal{I}_{x_S}) = min^K_{k=1} cosine\left(\theta(x^{db}_I), C_k\right) \tag{9}$$

where θ is the VGG-16 [42] function. We empirically observe that $K = 5$ gives the best results for retrieval. Other distance metrics typically used in clustering were considered but this gave the best results.

5.4 Experiments

We conduct an evaluation of the generative models on the proposed zero-shot setting and compare the results with those of existing methods in SBIR. We use the same metrics i.e Precision and mAP, for evaluation. We use the VGG-16 [42] model pre-trained on the Imagenet-1K dataset to obtain 4096 dimensional features for images. To extract the sketch features, we tune the network for sketch classification task using only the training sketches. We observed that this training gives only a marginal improvement in the performance and is hence optional.

Baselines Along with the state-of-the-art models for the SBIR task, we consider three popular algorithms [50] from the zero-shot image classification literature that do not explicitly use class label information and can be easily adopted to the zero-shot SBIR task. Let $(X_I, X_S) \in (\mathbb{R}^{N \times d_I}, \mathbb{R}^{N \times d_S})$ represent the image and sketch feature pairs from the training data respectively. We learn a mapping f from sketch features to image features, i.e $f : \mathbb{R}_I^d \to \mathbb{R}_S^d$ where d_I, d_S are the dimensions of the image and sketch vectors respectively. We describe these models below:

Direct Regression: The ZS-SBIR task is formulated as a simple regression problem, where each feature of the image feature vector is learnt from the sketch features. This is similar to the Direct Attribute prediction [23] which is a widely used baseline for zero-shot image classification.

Embarrassingly Simple Zero-Shot Learning: ESZSL was introduced by Romera-Paredes & Torr [35] as a method of learning bilinear compatibility matrix between images and attribute vectors in the context of zero-shot classification. In this work, we adapt the model to the ZS-SBIR task by mapping the sketch features to the image features using parallel training data from the train classes. The objective is to estimate $W \in \mathbb{R}^{d_S \times d_I}$ that minimizes the following loss:

$$||X_S W - X_I||_F^2 + \gamma \left|\left|X_I W^T\right|\right|_F^2 + \lambda \, ||X_S W||_F^2 + \beta \, ||W||_F^2 \tag{10}$$

where γ, λ, β are hyper-parameters.

Semantic Autoencoder: The Semantic Autoencoder (SAE) [21] proposes an autoencoder framework to encourage the re-constructibility of the sketch vector from the generated image vector. The loss term is given by:

$$||X_I - X_S W||_F^2 + \lambda \left|\left|X_I W^T - X_S\right|\right|_F^2 \tag{11}$$

We would like to note here that SAE, though simple, is currently the state-of-the-art among published models for zero-shot image classification task to the best of our knowledge.

Training We use Adam optimizer [19] with learning rate $\alpha = 2 \times 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a batch size of 64 and 128 for training the CVAE and CAAE respectively. We observe that the validation performance saturates at 25 epochs for the CVAE model and at 6000 iterations for the CAAE model. While training CAAE, we train the discriminator for 32 iterations for each training iteration of the encoder and decoder. We found that $N = 200$ i.e generating 200 image features for a given input sketch gives optimal performance and saturates afterwards. The reconstructibility parameter λ is set via cross-validation.

SAE has a single hyper-parameter and is solved using the Bartels-Stewart algorithm [2]. ESZSL has three hyper parameters γ, λ & β. We set $\beta = \gamma\lambda$ following the authors to get a closed form solution. We tune these hyper-parameters via a grid search from 10^{-6} to 10^7.

Table 3. The Precision and MAP evaluated on the retrieved 200 images in ZS-SBIR on the proposed split

Type	Evaluation Methods	Precision@200	mAP@200
SBIR methods	Baseline	0.106	0.054
	Siamese-1	0.243	0.134
	Siamese-2	0.251	0.149
	Coarse-grained triplet	0.169	0.083
	Fine-grained triplet	0.155	0.081
	DSH	0.153	0.059
ZSL methods	Direct Regression	0.066	0.022
	ESZSL	0.187	0.117
	SAE	0.238	0.136
Ours	CAAE	**0.260**	**0.156**
	CVAE	**0.333**	**0.225**

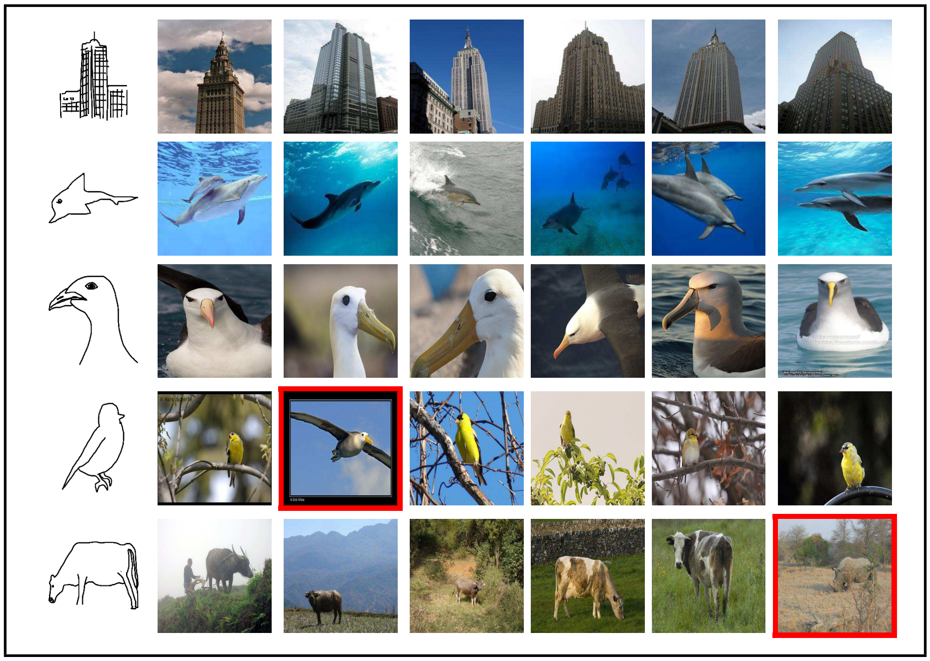

Fig. 3. Top 6 images retrieved for some input sketches using CVAE in the proposed zero-shot setting. Note that these sketch classes have never been encountered by the model during training. The red border indicates that the retrieved image does not belong to sketch's class. However, we would like to emphasize that the retrieved false positives do match the outline of the sketch

6 Results

The results of the evaluated methods for ZS-SBIR are summarized in Table 3. As observed in section-4.4, existing SBIR models perform poorly in the ZS-SBIR task. Both the proposed generative models out-perform the existing models indicating better latent alignment learning in the generative approach.

Qualitative Analysis: We show some of the retrieved images for sketch inputs of the unseen classes using the CVAE model in ZS-SBIR in Figure 3. We observe that the retrieved images closely match the outline of the sketch. We also observe that our model makes visually reasonable mistakes in the case of false positives wherein the retrieved images do have a significant similarity with the sketch even though they belong to a different class. For instance, in the last example the false positive that belongs to the class rhinoceros has a similar outline as that of the sketch. These may be considered not as an error but rather as a positive retrieval, but can only be evaluated qualitatively by an arduous manual task and may be attributed to data bias.

Human Evaluation: We aim to see how well the proposed zero-shot evaluation can substitute the fine-grained human evaluation. We randomly select 50 test sketches spanning all the unseen classes and then retrieve top 10 images per sketch from the database using the trained CVAE model. We compute the precision@10 for each of these sketches to get 50 such precision values (henceforth referred to as zero-shot scores).

Next, we present these sketch-image pairs to ten human evaluators. They were asked to evaluate each pair based on the outline, texture and overall shape associations, giving each pair a subjective score between 0 (no associations whatsoever) to 5 (perfect associations). We compute the average rating for the 10 retrieved images of each sketch and scale it down on a scale of 0-1. We compute the Pearson Correlation Coefficient (PCC) between the two scores across sketches, which was observed to be **0.65** indicating strong positive correlation between the two evaluation scores. The average human score across 50 sketches was observed to be **0.547**, whereas the average zero-shot score was **0.454**.

We repeat the above experiment using one of the baseline models, Coarse-grained Triplet Network. We observe a PCC of **0.69**. The average human score was **0.37** and the average zero-shot score was **0.238**. Across the two models studied, we observe that the scores are both high or both low, thus further strengthening the claim that methods working well on ZS-SBIR work well on fine-grained evaluation.

Feature visualization: To understand the kinds of features generated by the model, we visualize the generated image features of the test sketches in Figure 4 via the t-sne method. We make two observations, (i) the generated features are largely close to the true test image features (ii) multiple modalities of the distribution are captured by our model.

Performance Comparisons: Comparison among the current state-of-the-art models in the zero-shot setting of SBIR was already done in Section 4.4.

Direct regression from sketch to image feature space gives a precision value of 0.066. This serves as a baseline to evaluate other explicitly imposed regular-

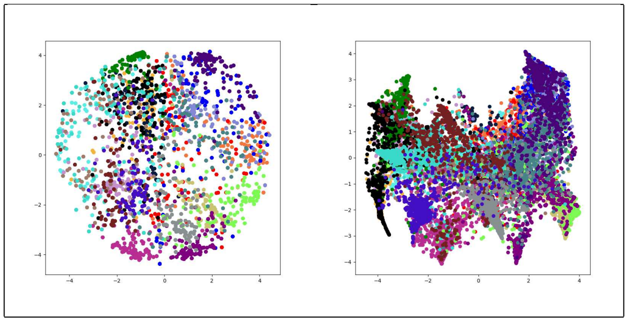

Fig. 4. T-SNE visualization of generated image features. Test data features are presented on the left and the predicted image features are on the right. Each color represents a particular class

izations in ESZSL and SAE. Our first observation is that the simple zero-shot learning models adapted to the ZS-SBIR task perform better than two state-of-the-art sketch based image retrieval models i.e Triplet network and DSH. SAE, which is the current state-of-the-art for zero-shot image classification, achieves the best performance among all the prior methods considered. SAE maps the sketches to images and hence generates a single image for a given sketch. This is similar to our proposed models except that our models generate a number of samples for a single sketch by filling the missing details from the latent distribution. Furthermore our model is non-linear whereas SAE is a simple linear projection. We believe that these generalizations over the SAE in our model leads to superior performance

Among the two models proposed, we observe that the CVAE models performs significantly better than the CAAE model. This may be attributed to the issue of instability while training adversarial models. We observe that the training error of the CVAE models is much more smoother compared to the CAAE model. We observe that using the reconstruction loss leads to a 3% improvement on the precision.

7 Conclusion

We identified major drawbacks in current evaluation schemes in sketch-based image retrieval (SBIR) task. To this end, we pose the problem of sketch-based retrieval in a zero-shot evaluation framework (ZS-SBIR). By making a careful split in the "Sketchy" dataset, we provide a benchmark for this task. We then evaluate current state-of-the-art SBIR models in this framework and show that the performance of these models drop significantly, thus exposing the class-specific learning which is inherent to these models. We then pose the SBIR problem as a generative task and propose two conditional generative models which achieve significant improvement over the existing methods in ZS-SBIR setting.

References

1. Akata, Z., Reed, S.E., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: CVPR. pp. 2927–2936. IEEE Computer Society (2015), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#AkataRWLS15 4
2. Bartels, R.H., Stewart, G.W.: Solution of the matrix equation $AX + XB = C$. Comm. ACM **15**, 820–826 (1972) 11
3. Cao, X., Zhang, H., Liu, S., Guo, X., Lin, L.: Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor. In: ICCV. pp. 313–320. IEEE Computer Society (2013), http://dblp.uni-trier.de/db/conf/iccv/iccv2013.html#CaoZLGL13 2
4. Cao, Y., Wang, C., Zhang, L., Zhang, L.: Edgel index for large-scale sketch-based image search. In: CVPR. pp. 761–768. IEEE Computer Society (2011), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2011.html#CaoWZZ11 2
5. Cao, Y., Wang, H., Wang, C., Li, Z., Zhang, L., Zhang, L.: Mindfinder: interactive sketch-based image search on millions of images. In: Bimbo, A.D., Chang, S.F., Smeulders, A.W.M. (eds.) ACM Multimedia. pp. 1605–1608. ACM (2010), http://dblp.uni-trier.de/db/conf/mm/mm2010.html#CaoWWLZZ10 2
6. Chidambaram, M., Qi, Y.: Style transfer generative adversarial networks: Learning to play chess differently. CoRR **abs/1702.06762** (2017), http://dblp.uni-trier.de/db/journals/corr/corr1702.html#ChidambaramQ17 8
7. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 539–546. IEEE (2005) 2, 4, 6, 7
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009) 5, 7
9. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE Trans. Vis. Comput. Graph. **17**(11), 1624–1636 (2011), http://dblp.uni-trier.de/db/journals/tvcg/tvcg17.html#EitzHBA11 2
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS. pp. 2121–2129 (2013), http://dblp.uni-trier.de/db/conf/nips/nips2013.html#FromeCSBDRM13 4
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) NIPS. pp. 2672–2680 (2014), http://dblp.uni-trier.de/db/conf/nips/nips2014.html#GoodfellowPMXWOCB14 10
12. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2). pp. 1735–1742. IEEE Computer Society (2006), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2006-2.html#HadsellCL06 6
13. Hu, R., Collomosse, J.P.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Computer Vision and Image Understanding **117**(7), 790–806 (2013), http://dblp.uni-trier.de/db/journals/cviu/cviu117.html#HuC13 2
14. Hu, R., Collomosse, J.P.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Computer Vision and Image Understanding

117(7), 790–806 (2013), http://dblp.uni-trier.de/db/journals/cviu/cviu117.html#HuC13 4

15. Hu, R., Wang, T., Collomosse, J.P.: A bag-of-regions approach to sketch-based image retrieval. In: Macq, B., Schelkens, P. (eds.) ICIP. pp. 3661–3664. IEEE (2011), http://dblp.uni-trier.de/db/conf/icip/icip2011.html#HuWC11 2
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arxiv (2016) 8
17. James, S., Fonseca, M.J., Collomosse, J.P.: Reenact: Sketch based choreographic design from archival dance footage. In: Kankanhalli, M.S., Rueger, S., Manmatha, R., Jose, J.M., van Rijsbergen, K. (eds.) ICMR. p. 313. ACM (2014), http://dblp.uni-trier.de/db/conf/mir/icmr2014.html#JamesFC14 2
18. John Eakins, M.G.: Content-based image retrieval 2
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14 11
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 8, 9
21. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. CoRR **abs/1704.08345** (2017), http://dblp.uni-trier.de/db/journals/corr/corr1704.html#KodirovXG17 4, 11
22. Kumar Verma, V., Arora, G., Mishra, A., Rai, P.: Generalized zero-shot learning via synthesized examples. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 4
23. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 453–465 (2014), http://dblp.uni-trier.de/db/journals/pami/pami36.html#LampertNH14 4, 5, 11
24. Li, K., Pang, K., Song, Y.Z., Hospedales, T.M., Zhang, H., Hu, Y.: Fine-grained sketch-based image retrieval: The role of part-aware attributes. In: WACV. pp. 1–9. IEEE Computer Society (2016), http://dblp.uni-trier.de/db/conf/wacv/wacv2016.html#LiPSHZH16 2
25. Li, X., Liao, S., Lan, W., Du, X., Yang, G.: Zero-shot image tagging by hierarchical semantic embedding. In: Baeza-Yates, R.A., Lalmas, M., Moffat, A., Ribeiro-Neto, B.A. (eds.) SIGIR. pp. 879–882. ACM (2015), http://dblp.uni-trier.de/db/conf/sigir/sigir2015.html#LiLLDY15 4
26. Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L.: Deep sketch hashing: Fast free-hand sketch-based image retrieval. CoRR **abs/1703.05605** (2017), http://dblp.uni-trier.de/db/journals/corr/corr1703.html#LiuSSLS17 2, 4, 5, 7, 8
27. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I.: Adversarial autoencoders. In: International Conference on Learning Representations (2016), http://arxiv.org/abs/1511.05644 8, 10
28. Mishra, A., Reddy, M., Mittal, A., Murthy, H.A.: A generative model for zero shot learning using conditional variational autoencoders. arXiv preprint arXiv:1709.00663 (2017) 4
29. Mishra, A., Verma, V.K., Reddy, M., Rai, P., Mittal, A., et al.: A generative approach to zero-shot and few-shot action recognition. arXiv preprint arXiv:1801.09086 (2018) 4
30. Parui, S., Mittal, A.: Similarity-invariant sketch-based image retrieval in large databases. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV (6). Lecture Notes in Computer Science, vol. 8694, pp. 398–414. Springer (2014), http://dblp.uni-trier.de/db/conf/eccv/eccv2014-6.html#ParuiM14 2, 3

31. Qi, Y., Song, Y.Z., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. In: ICIP. pp. 2460–2464. IEEE (2016), http://dblp.uni-trier.de/db/conf/icip/icip2016.html#QiSZL16 2
32. Qi, Y., Song, Y.Z., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. In: ICIP. pp. 2460–2464. IEEE (2016), http://dblp.uni-trier.de/db/conf/icip/icip2016.html#QiSZL16 6
33. Ramakrishnan, S.K., Pal, A., Sharma, G., Mittal, A.: An empirical evaluation of visual question answering for novel objects. CoRR **abs/1704.02516** (2017), http://dblp.uni-trier.de/db/journals/corr/corr1704.html#RamakrishnanPSM17 4
34. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text-to-image synthesis. In: Proceedings of The 33rd International Conference on Machine Learning (2016) 8
35. Romera-Paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. In: Bach, F.R., Blei, D.M. (eds.) ICML. JMLR Workshop and Conference Proceedings, vol. 37, pp. 2152–2161. JMLR.org (2015), http://dblp.uni-trier.de/db/conf/icml/icml2015.html#Romera-ParedesT15 4, 11
36. Saavedra, J.M.: Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In: ICIP. pp. 2998–3002. IEEE (2014), http://dblp.uni-trier.de/db/conf/icip/icip2014.html#Saavedra14 2
37. Saavedra, J.M.: Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In: ICIP. pp. 2998–3002. IEEE (2014), http://dblp.uni-trier.de/db/conf/icip/icip2014.html#Saavedra14 4
38. Saavedra, J.M., Barrios, J.M.: Sketch based image retrieval using learned keyshapes (lks). In: Xie, X., Jones, M.W., Tam, G.K.L. (eds.) BMVC. pp. 164.1–164.11. BMVA Press (2015), http://dblp.uni-trier.de/db/conf/bmvc/bmvc2015.html#SaavedraB15 2
39. Saavedra, J.M., Barrios, J.M.: Sketch based image retrieval using learned keyshapes (lks). In: Xie, X., Jones, M.W., Tam, G.K.L. (eds.) BMVC. pp. 164.1–164.11. BMVA Press (2015), http://dblp.uni-trier.de/db/conf/bmvc/bmvc2015.html#SaavedraB15 4
40. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans. Graph. **35**(4), 119 (2016), http://dblp.uni-trier.de/db/journals/tog/tog35.html#SangkloyBHH16 2, 3, 4, 5, 6, 7
41. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823. IEEE Computer Society (2015), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#SchroffKP15 6, 8
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014) 7, 10
43. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS. pp. 935–943 (2013), http://dblp.uni-trier.de/db/conf/nips/nips2013.html#SocherGMN13 4
44. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) NIPS. pp. 3483–3491 (2015), http://dblp.uni-trier.de/db/conf/nips/nips2015.html#SohnLY15 8, 9
45. Teney, D., van den Hengel, A.: Zero-shot visual question answering. CoRR **abs/1611.05546** (2016), http://dblp.uni-trier.de/db/journals/corr/corr1611.html#TeneyH16a 4

46. Verma, V.K., Rai, P.: A simple exponential family framework for zero-shot learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 792–808. Springer (2017) 4
47. Wang, F., Kang, L., Li, Y.: Sketch-based 3d shape retrieval using convolutional neural networks. CoRR **abs/1504.03504** (2015), http://dblp.uni-trier.de/db/journals/corr/corr1504.html#WangKL15 2
48. Wang, W., Pu, Y., Verma, V.K., Fan, K., Zhang, Y., Chen, C., Rai, P., Carin, L.: Zero-shot learning via class-conditioned deep generative models. arXiv preprint arXiv:1711.05820 (2017) 4
49. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. CoRR **abs/1603.08895** (2016), http://dblp.uni-trier.de/db/journals/corr/corr1603.html#XianA0N0S16 4
50. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: IEEE Computer Vision and Pattern Recognition (CVPR) (2017) 4, 5, 11
51. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: CVPR. pp. 799–807. IEEE Computer Society (2016), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2016.html#YuLSXHL16 2, 3, 4
52. Yu, Q., Yang, Y., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-net: A deep neural network that beats humans. International Journal of Computer Vision **122**(3), 411–425 (2017), http://dblp.uni-trier.de/db/journals/ijcv/ijcv122.html#YuYLSXH17 4
53. Zhang, Y., Gong, B., Shah, M.: Fast zero-shot image tagging. CoRR **abs/1605.09759** (2016), http://dblp.uni-trier.de/db/journals/corr/corr1605.html#ZhangGS16 4
54. Zhou, R., Chen, L., Zhang, L.: Sketch-based image retrieval on a large scale database. In: Babaguchi, N., Aizawa, K., Smith, J.R., Satoh, S., Plagemann, T., Hua, X.S., Yan, R. (eds.) ACM Multimedia. pp. 973–976. ACM (2012), http://dblp.uni-trier.de/db/conf/mm/mm2012.html#ZhouCZ12 2