Adversarial Joint-Distribution Learning for Novel Class Sketch-Based Image Retrieval

Anubha Pandey[1] anubhap93@gmail.com
Ashish Mishra[1] mishra@cse.iitm.ac.in
Vinay Kumar Verma[2] vkverma@cse.iitk.ac.in
Anurag Mittal[1] amittal@cse.iitm.ac.in
[1]Indian Institute of Technology Madras [2]Indian Institute of Technology Kanpur

Abstract

In the information retrieval task, sketch-based image retrieval (SBIR) has drawn significant attention owing to the ease with which sketches can be drawn. The existing deep learning methods for the SBIR are very unrealistic in the real scenario, and its performance reduces drastically for unseen class test examples. Recently, Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) [37, 47] has drawn a lot of attention due to its ability to retrieve the novel/unseen class images at test time. These methods try to project sketch features into the image domain by learning a distribution conditioned on the sketch. We propose a new framework for ZS-SBIR that models joint distribution between the sketch and image domain using a generative adversarial network [29]. The joint distribution modeling ability of our generative model helps to reduce the domain gap between the sketches and images. Our framework helps to synthesize the novel class image features using sketch features. The generative ability of our model for the unseen/novel classes, conditioned on sketch feature, allows it to perform well on the seen as well as unseen class sketches. We conduct extensive experiments on two widely used SBIR benchmark datasets- Sketchy and Tu-Berlin and obtain significant improvement over the existing state-of-the-art. We will release the code publicly for reproducibility of results.

1. Introduction

Content-based image retrieval techniques[24] retrieve images from a huge database by analyzing the contents of an image like color, texture, shape, etc. Image-based queries give better results as it captures the contents i.e., color, texture, shape, etc. better. However, it is not practically possible to always find an appropriate image query similar to the image to be retrieved. Hence, sketches are gaining more popularity in content-based image retrieval(CBIR) over images. Free-hand drawn sketches are convenient to draw as shapes of the image can be remembered easily. The Sketch-based Image Retrieval(SBIR) task refers to the retrieval of images from a huge database based on sketch queries [3, 4, 48, 30]. Sketch-based recognition tasks such as forensic face photo-sketch recognition [9] and fine grain sketch-based image retrieval (FG-SBIR) [2, 39, 49] are becoming widespread. The development of digital touch-screen (Smart-phones, Ipads, and Wacom, etc.) devices owes to the popularity of hand-drawn sketches. SBIR remains a challenging and difficult problem due to a

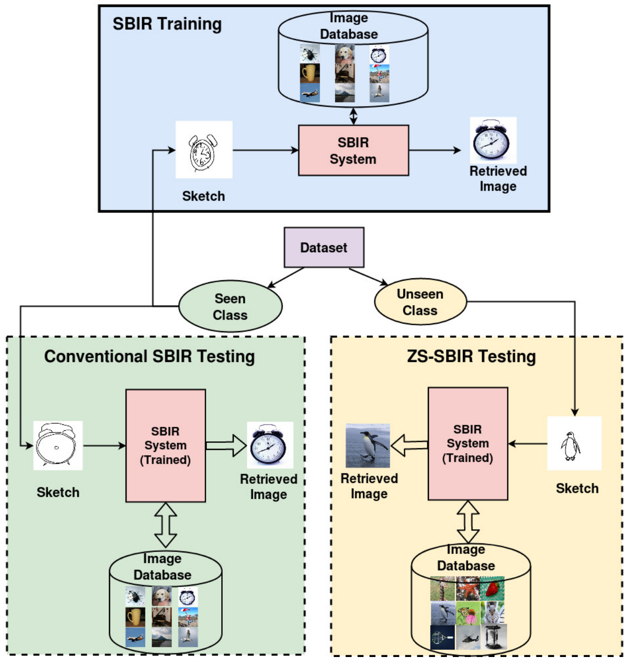

Figure 1. An Overview of Conventional and Zero-Shot SBIR approaches: (a)In Conventional SBIR input query during testing comes from seen classes (b)In Zero-shot SBIR input query during testing comes from unseen classes

significant inconsistency between the sketch domain and the image domain. Free-hand drawn sketches are ambiguous and vary significantly from person to person based on the attributes they want to emphasize. To reduce the domain-gap between the sketch and image domain, recent methods [7, 13, 14] use cross-domain transfer learning and knowledge transfer learning. They project them into a common subspace and further use the projected features for the task of image retrieval. This kind of approach learns to embed within each domain, and therefore can not generalize well for test data that have a significant variance with that of the training examples.

However, existing methods in SBIR assume that at all the sketch query at test time come from seen classes i.e., from the classes on which network is trained. Hence these methods give poor performance when the query sketch comes from an unseen class. The main reasons as to why traditional SBIR methods perform poorly on an unseen class are as follows: 1) Most of the approaches are naturally biased towards predicting the seen classes, and 2) They do not leverage any transfer of information from seen classes to unseen classes. This necessitates posing the problem in the zero-shot learning setting. Figure 1 shows the outline of the proposed approach for ZS-SBIR, which also differentiates between conventional SBIR and ZS-SBIR model. In this paper, we propose a method for the task of Sketch-based Image Retrieval in zero-shot set-up(ZS-SBIR) to handle an unseen/novel sketch class at the test time. Recently, researchers have explored techniques of Zero-shot Learning in image classification problems [42, 31, 25, 44, 52, 17].

There are a few recent works [47, 37, 5] proposed on ZS-SBIR. ZSIH [37] and Doodle to search [5] use sketch class description as side information along with sketches to train the model, whereas [47] proposed two models without using side information conditional variational autoencoder(CVAE) based and conditional adversarial autoencoder(CAAE) based model. Due to the explosive growth of new categories, it is not practically possible to get class descriptions for every new class.

We use a generative model based on the joint-adversarial learning [29] for the ZS-SBIR problem. We learn joint distribution between sketch and image domains using a generative model and synthesize novel domain examples conditioned on the other. The joint distribution modeling ability of our generative model helps to reduce the domain gap between the sketches and images. Our model synthesizes the image features condition on sketch features without using any side information. Below are the contributions of this paper:

- We propose to use a new framework for ZS-SBIR problem that learns a joint distribution between image features and sketch features using jointGAN [29].
- We propose to use a Maximum Mean Discrepancy (MMD)loss [11] in jointGAN [29] to quantify the distance between the means of the two different class distributions. It can be used to distinguish between pairs of real and generated features of both images and sketches which belong to different classes.
- Our method yields significantly better results in the zero-shot setting without using any side information (e.g., word2vec representation of the class attributes[21, 28]) on both Sketchy and Berlin datasets.

2. Related work

Traditional methods in SBIR represent sketches and edge-maps of images using hand-crafted features like Gradient Flow HOG descriptor [14], The Learned Key Shape(LKS) [33], Histogram Edge Orientation(HELO) [34] etc. and match these features for the task of image retrieval. Sarthak et al.[27] proposed a similarity invariant chain descriptors to represent sketches and edge-map of images and used a dynamic programming-based algorithm for further retrieval. With the advancement in deep learning techniques, researchers started using features from deep networks [50, 35, 47, 37] for a better representation of sketches and images for retrieval. These deep features are invariant to deformation, color, and texture. However, the cross-domain discrepancy between sketches and images can't be well remedied by the traditional SBIR methods. It is difficult to match edge-maps to corresponding sketches with larger variations and ambiguity. Recent techniques in SBIR project sketches and images into a common space to reduce the domain gap. [30, 50] used Siamese architecture, [35] used triplet ranking loss for coarse-grained SBIR and [19] proposed a deep architecture for extracting the binary codes from the sketches and the images. Other approaches project sketch to image domain or vice-versa [8, 15] such that sketches and images of the same class become close to each other and of different classes are separated by a margin.

Conventional SBIR methods assume the availability of training samples from all the classes. Hence give poor performance when data comes from unseen classes during test time. Zero-Shot Learning (ZSL) techniques are capable of handling unseen/novel class at test time and has drawn significant attention. ZSL use class attributes/descriptions to learn the interaction between visual space and semantic space. There exist two standard approaches for the ZSL: (1) Embedding-based ZSL (2) Synthesis-bases ZSL. Embedding-based ZSL is further categorized based on the direction of the embedding function. One approach learns the mapping from visual to semantic space and vice-versa [46, 1, 26, 43, 41, 23, 22]. The other approach learns the bilinear embedding between both the visual and semantic space [42]. It maps the visual features, and semantic attributes to a common subspace such that those belong to the same class are closer and of different classes are separated from each other. Synthesis-based ZSL [18, 45, 46] uses generative models to synthesize unseen class features based on the semantic attributes hence reduce the ZSL problem to supervised image classification problem. Generative models are used to learn the underlying image distribution conditioned on some semantic attributes.

Recently, Shen et al.[37], Dey et al.[5] and Yelamarthi et al.[47] proposed SBIR in Zero-Shot framework. Shen et al.[37] in their proposed ZSIH approach, combined zero-shot learning and sketch-based image retrieval using a cross-modal hashing scheme. Dey et al.[5] proposed a ZS-SBIR framework that learns a common embedding space for both the sketch and image domains. Both these methods [37, 5] used sketch class descriptions[28] as side information along with sketch features for establishing the semantic relationship between the image feature space and sketch feature space. In contrast, Yelamarthi et al.[47] proposed two similar autoencoder-based generative models, CAAE(Conditional Adversarial Autoencoder) [20] and CVAE(Conditional Variational Autoencoder)[38] for zero-shot SBIR without using any side information.

3. Proposed Approach

3.1. Zero-Shot SBIR (ZS-SBIR)

In the zero-shot setting, we divide the dataset into two disjoint sets of seen(S) and unseen(U) classes of sketches. During training pairs of sketch and images from seen classes are available whereas, during testing, only sketches from unseen classes are present.

Let $A = \{(\mathbf{x_i}^{skt}, \mathbf{x_i}^{img}, l_i) | l_i \in \mathcal{L}\}$ be the triplet of sketch, image, and the class label where $\mathcal{L}$ is the set of all class labels. We partition the class labels in the dataset into $\mathcal{L}_{train}$ and $\mathcal{L}_{test}$ for the train and test respectively. Let $A_{tr} = \{\mathbf{x_i}^{skt}, \mathbf{x_i}^{img}, l_i | l_i \in L_{train}\}$ and $A_{te} = \{\mathbf{x_i}^{skt}, \mathbf{x_i}^{img}, l_i | l_i \in L_{test}\}$ be the partition of A into train and test sets. For simplicity, Let's represent $\mathbf{x}^{skt}$ as ”$\mathbf{y}$” and $\mathbf{x}^{img}$ as ”$\mathbf{x}$” throughout this paper.

3.2. Adversarial Joint Distribution Learning

In this section, we describe the approach to learn the joint distribution of two random variables, in our case image features and sketch(attribute) features. Suppose $q(\mathbf{x})$ and $q(\mathbf{y})$ are the marginal distributions of two random variables $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ respectively. In our proposed approach $\mathbf{X}$ and $\mathbf{Y}$ represent image features and sketch features(attributes) respectively. Attribute for each image is its corresponding sketch feature. Usually, the true distribution $q(\mathbf{x})$ and $q(\mathbf{y})$ are not known, whereas samples $\{\mathbf{x_i}\}_{i=1}^{N}$ and $\{\mathbf{y_i}\}_{i=1}^{N}$ from the both distributions are available. The joint distribution of image $\mathbf{x}$ and attribute $\mathbf{y}$ can be represented as a product of the marginal and a conditional in two ways: $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})q(\mathbf{y}|\mathbf{x}) = q(\mathbf{y})q(\mathbf{x}|\mathbf{y})$.One feature can be synthesized given the other feature using the conditional distributions $q(\mathbf{x}|\mathbf{y})$ and $q(\mathbf{y}|\mathbf{x})$.

Using CGAN, we can learn the joint distribution between two random variables by either assuming we have the marginal distribution or learn the marginal distribution of one random variable using conventional GANs and then learn conditional distribution on top of it. Since there is no flow of information between marginals and conditionals during training, this is not a proper method to determine the joint distribution.

To address this issue, we learn the joint distribution of images and sketches using a method mentioned in JointGAN [29]. We use a combination of two generators to determine the marginal and conditional distribution simultaneously along with a single discriminator (or critic) for training, as shown in Figure 2. The formulation of jointGAN is given as:

$$\hat{\mathbf{x}} = G_\alpha(\mathbf{z_1}), \quad \hat{\mathbf{y}} = G_\phi(\hat{\mathbf{x}}, \mathbf{z_2}) \tag{1}$$

$$\hat{\mathbf{y}} = G_\beta(\mathbf{z_3}), \quad \hat{\mathbf{x}} = G_\theta(\hat{\mathbf{y}}, \mathbf{z_4}) \tag{2}$$

Where $G_\alpha(.)$, $G_\beta(.)$ are marginal generators of images and sketches(attributes) respectively and $G_\theta(.)$ and $G_\phi(.)$ are conditional generators of images and sketches(attributes) respectively. $\mathbf{z_1}$, $\mathbf{z_2}$, $\mathbf{z_3}$,and $\mathbf{z_4}$ are independent noise sampled from unit Gaussian. Both the generators $G_\alpha(.)$ and $G_\theta(.)$ synthesize image features, the only difference is that $G_\theta(.)$ takes sketch(attribute) $\mathbf{y}$ as a conditional input. If we replace $\mathbf{y}$ with $\mathbf{0}$ vector of same dimension in G_θ then it is the same as $G_\alpha(.)$. Therefore, we can couple the parameters α and θ together. Similarly β and ϕ can also be coupled together. $G_\alpha(.)$ and $G_\beta(.)$ can be represented as :

$$G_\alpha(.) = G_\theta(\mathbf{0}, .), G_\beta(.) = G_\phi(\mathbf{0}, .) \tag{3}$$

Now, let $p_\theta(\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{y})$ be the marginal and conditional distribution learned for images using generator G_θ. Similarly, let $p_\phi(\mathbf{y})$ and $p_\phi(\mathbf{y}|\mathbf{x})$ be the marginal and conditional distribution learned for sketches(attributes) using generator G_ϕ. The possible combinations of all the marginal and conditional distributions (learned using both the generators G_ϕ and G_θ) to determine joint distribution of images and sketches(attributes) are shown below:

$$p_1(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})p_\phi(\mathbf{y}|\mathbf{x}), \;\; p_2(\mathbf{x}, \mathbf{y}) = q(\mathbf{y})p_\theta(\mathbf{x}|\mathbf{y}) \tag{4}$$

$$p_3(\mathbf{x}, \mathbf{y}) = p_\alpha(x)p_\phi(\mathbf{y}|\mathbf{x}), \;\; p_4(\mathbf{x}, \mathbf{y}) = p_\beta(\mathbf{y})p_\theta(\mathbf{x}|\mathbf{y}) \tag{5}$$

where $q(\mathbf{x})$ and $q(\mathbf{y})$ are true distributions of images and sketches(attributes) respectively.

For adversarial learning, we use four Discriminators(binary critics) corresponding to each joint distributions

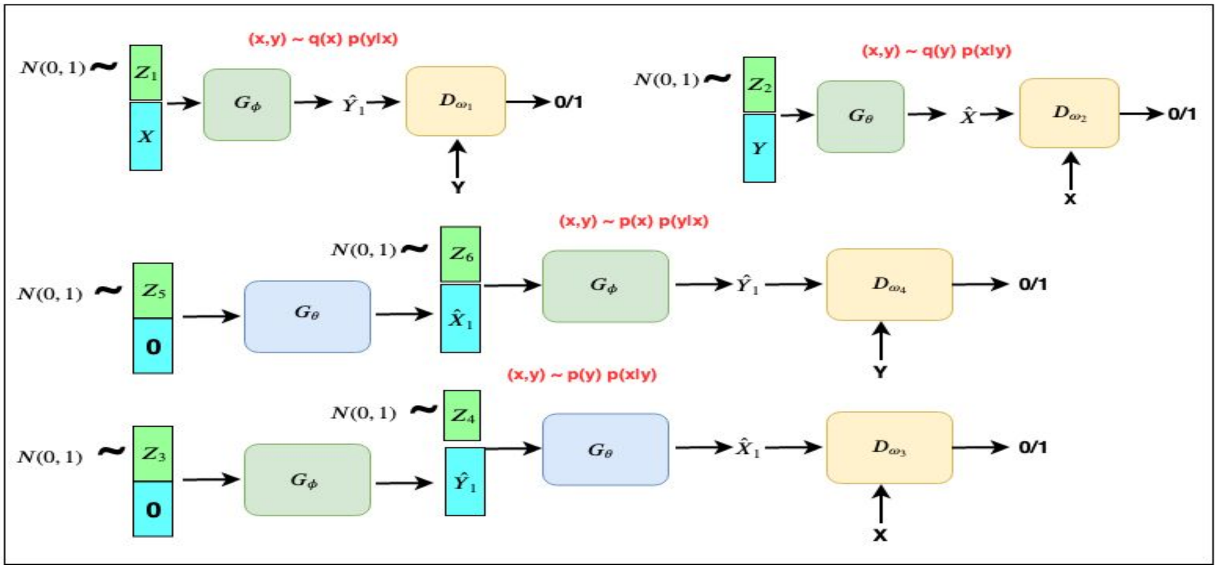

Figure 2. An illustration of determining the joint distribution of images and sketches(attributes) from the possible combinations of all the marginal and conditional distributions(learned using both the generators G_ϕ and G_θ). We simultaneously learn marginal and conditional distributions of Images and Attributes.

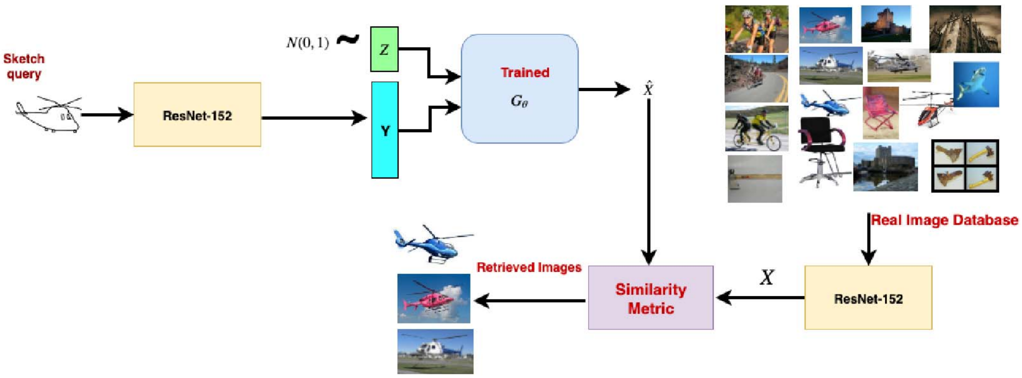

Figure 3. An illustration of image retrieval process of our proposed model.

$p_1(\mathbf{x},\mathbf{y})$, $p_2(\mathbf{x},\mathbf{y})$, $p_3(\mathbf{x},\mathbf{y})$ and $p_4(\mathbf{x},\mathbf{y})$, to mimic a four-class classifier. Each Discriminator distinguishes between paired samples of images and sketches(attributes) taken from the corresponding learned joint distributions and the true joint distribution $q(\mathbf{x},\mathbf{y})$. Here we have paired samples of images and sketches(attributes) in our training dataset. Let the discriminator D_{ω_i} corresponds to the $p_i(\mathbf{x},\mathbf{y})$ joint distribution, where $D_{\omega_i} \in (0,1)$. The minimax objective for jointGAN is:

$$\min_{\theta,\phi} \max_{\omega} L_{jGAN}(\theta,\phi,\omega) = \sum_{i=1}^{i=4} E_{p_i(\mathbf{x},\mathbf{y})}[logD_{\omega_i}(\mathbf{x},\mathbf{y})] \quad (6)$$

Where $\omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ is list of all parameter of the discriminators. The equilibrium of this objective is achieved if and only if $p_1(\mathbf{x},\mathbf{y}) = p_2(\mathbf{x},\mathbf{y}) = p_3(\mathbf{x},\mathbf{y}) = p_4(\mathbf{x},\mathbf{y})$.

In Equation 6 expectations $E_{p_3(\mathbf{x},\mathbf{y})}(.)$ and $E_{p_4(\mathbf{x},\mathbf{y})}(.)$ are approximated with purely synthesized joint samples, whereas $E_{p_1(\mathbf{x},\mathbf{y})}(.)$ and $E_{p_2(\mathbf{x},\mathbf{y})}(.)$ are approximated with conditionally synthesized sample given samples from the true marginals. To train our model to generate more discriminative features we further use Cycle Consistency loss [40, 53] and MMD loss [11] as a regularizer. Both the losses are explained below:

Cycle Consistency loss

To regularize the model, we use the constraint of cycle-consistency. Cycle consistency loss ensures we generate image samples similar to the original samples using a series of learned distributions, as shown below:

$q(\mathbf{x}) \rightarrow \mathbf{x} \rightarrow p_\phi(\mathbf{y}|\mathbf{x}) \rightarrow \mathbf{y} \rightarrow p_\theta(\mathbf{x}|\mathbf{y}) \rightarrow \hat{x}$.

$\hat{\mathbf{x}}$ should be very similar to real $\mathbf{x}$, this implies that $||\mathbf{x} - \hat{\mathbf{x}}||$ approximate to zero.

Similarly, attributes samples similar to that of the original attributes are generated as shown below:

$q(\mathbf{y}) \rightarrow \mathbf{y} \rightarrow p_\theta(\mathbf{x}|\mathbf{y}) \rightarrow \mathbf{x} \rightarrow p_\phi(\mathbf{y}|\mathbf{x}) \rightarrow \hat{\mathbf{y}}$.

$\hat{\mathbf{y}}$ also should be very similar to real $\mathbf{y}$. The cycle-consistency loss is defined as :

$$C_{\theta,\phi}(\mathbf{x},\mathbf{y}) = E_{\mathbf{x}\sim q(\mathbf{x}),\mathbf{y}\sim p_\phi(\mathbf{y}|\mathbf{x}),\hat{\mathbf{x}}\sim p_\theta(\mathbf{x}|\mathbf{y})}||\mathbf{x}-\hat{\mathbf{x}}|| + E_{\mathbf{y}\sim q(\mathbf{y}),\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{y}),\hat{\mathbf{y}}\sim p_\phi(\mathbf{y}|\mathbf{x})}||\mathbf{y}-\hat{\mathbf{y}}|| \quad (7)$$

The objective function becomes:

$$\min_{\theta,\phi}\max_{\omega} L_{jGAN}(\theta,\phi,\omega) = \sum_{i=1}^{i=4} E_{p_i(\mathbf{x},\mathbf{y})}[logD_{\omega i}(\mathbf{x},\mathbf{y})] + \lambda_1 * C_{\theta,\phi}(\mathbf{x},\mathbf{y}) \quad (8)$$

where $C_{\theta,\phi}(\mathbf{x},\mathbf{y})$ is a cycle consistency regularization term, $\omega = \{\omega_1,\omega_2,\omega_3,\omega_4\}$ and λ_1 is a hyperparameter.

Maximum Mean Discrepancy loss

We propose to use a Maximum Mean Discrepancy (MMD)loss [11] in jointGAN [29] to quantify the distance between the mean of the two different class distributions. It can be used to distinguish between pairs of real and generated features of both images and sketches which belong to different classes. The MMD loss is a kernel-based distance function between pairs of synthesized and real samples. Here we compute MMD loss between generated image features $\hat{\mathbf{x}}$ and real image features $\mathbf{x}$. Similarly, for generated sketches features $\hat{\mathbf{y}}$ and real sketch features $\mathbf{y}$. The loss function is given by:

$$L_{Img}^{mmd}(\mathbf{x},\hat{\mathbf{x}}) = \sum_{j=1}^{j=N}\sum_{j'=1}^{j'=N} k(\mathbf{x_j},\mathbf{x_{j'}}) - 2\sum_{j=1}^{j=N}\sum_{i=1}^{i=N} k(\mathbf{x_j},\hat{\mathbf{x}}_\mathbf{i}) + \sum_{i=1}^{i=N}\sum_{i'=1}^{i'=N} k(\hat{\mathbf{x}}_\mathbf{i},\hat{\mathbf{x}}_\mathbf{i'}) \quad (9)$$

We leverage the linear combination of multiple RBF kernels ($k(\mathbf{x},\hat{\mathbf{x}})$) that is defined as :

$$k(\mathbf{x},\hat{\mathbf{x}}) = \sum_n \eta_n \exp\left(\frac{-||\mathbf{x}-\hat{\mathbf{x}}||^2}{2\sigma_n}\right) \quad (10)$$

where σ_n is the standard deviation and η_n is the weight factor for n^{th} RBF kernel. Similarly, we can define for sketch(attribute) $\mathbf{y}$:

$$L_{Att}^{mmd}(\mathbf{y},\hat{\mathbf{y}}) = \sum_{j=1}^{j=N}\sum_{j'=1}^{j'=N} k(\mathbf{y_j},\mathbf{y_{j'}}) - 2\sum_{j=1}^{j=N}\sum_{i=1}^{i=N} k(\mathbf{y_j},\hat{\mathbf{y}}_\mathbf{i}) + \sum_{i=1}^{i=N}\sum_{i'=1}^{i'=N} k(\hat{\mathbf{y}}_\mathbf{i},\hat{\mathbf{y}}_\mathbf{i'}) \quad (11)$$

The overall MMD loss is defined as the sum of Eq. 9 and Eq. 11:

$$L_{mmd} = L_{Img}^{mmd} + L_{Att}^{mmd} \quad (12)$$

Now, the overall objective function for our proposed approach is defined as :

$$\min_{\theta,\phi}\max_{\omega} L_{jGAN}(\theta,\phi,\omega) = \sum_{i=1}^{i=4} E_{p_i(\mathbf{x},\mathbf{y})}[logD_{\omega_i}(\mathbf{x},\mathbf{y})] + \lambda_1 * C_{\theta,\phi}(\mathbf{x},\mathbf{y}) + \lambda_2 * L_{mmd} \quad (13)$$

Where hyper-parameters λ_1 and λ_2 corresponds to cycle consistency loss and MMD loss respectively.

3.3. Image retrieval for unseen class sketch query

The image retrieval process from real image database for unseen class sketches is illustrated in Figure 3.

1. Obtain sketch class attributes $\mathbf{y}$ by extracting features for sketch query using pre-trained ResNet-152.

2. Pass random noise $\mathbf{z}$ and attribute $\mathbf{y}$ to the trained generator G_θ which generates image feature $\hat{\mathbf{x}}$ corresponding to the class attribute $\mathbf{y}$ as $\hat{\mathbf{x}} = G_\theta(\mathbf{z},\mathbf{y})$.

3. Find the similarity between generated image feature $\hat{\mathbf{x}}$ and the image feature $\mathbf{x}$ from the real image database and retrieve top-K similar images from the real image database.

4. Implementation and Results

4.1. Dataset and Visual Feature

For the evaluation of our proposed model, we perform experiments on two challenging datasets- Sketchy [35] and TU-Berlin [6]. There are 125 sketch classes in the Sketchy [35] dataset with 75471 hand-drawn sketches. Initially, Sketchy dataset contains 12500 real images corresponding to the sketches. [19] extend the original Sketchy dataset by introducing 60502 more real images from 125 different categories. TU-Berlin([6]) extended introduced by [19, 51] has 20000 sketches and 204489 images from 250 classes and is a large scale dataset.

We use visual features of sketches as conditioning attributes in our proposed generative model. We extract the visual features of images and sketches from the last fully connected layer of ResNet-152 [12] pre-trained on ImageNet-1000 dataset without any fine-tuning. We extract 2048-dimensional feature vectors. We believe that fine-tuning on this dataset will result in better performance of our model. For a fair comparison with our proposed model, we reproduce the result in ResNet-152 features for all baseline models.

Type	Method	Sketchy Dataset		TU Berlin Dataset	
		Precision@200	mAP@200	Precision@200	mAP@200
SBIR	Baseline	0.176	0.099	0.139	0.083
	Siamese-1 [30]	0.243	0.134	0.127	0.061
	Siamese-2 [50]	0.251	0.149	0.133	0.067
	Fine-Grained Triplet [35]	0.155	0.081	0.086	0.050
	Coarse-Grained Triplet [36]	0.169	0.083	0.128	0.057
Zero-Shot	Direct Regression	0.066	0.022	0.117	0.062
	ESZSL[32]	0.187	0.117	0.131	0.072
	DAP [18]	0.078	0.071	0.075	0.067
	SAE [16]	0.238	0.136	0.152	0.084
	CAAE [47]	0.240	0.146	0.159	0.094
	CVAE [47]	0.269	0.159	0.182	0.109
	Ours	**0.319**	**0.221**	**0.204**	**0.129**

Table 1. Precision@200 and mAP@200 results on the traditional SBIR methods and ZSL methods in the ZS-SBIR setup. Note that we re-implement all state-of-the-art methods for fair comparison. [47] proposed two models CAAE and CVAE.

Sketchy Dataset (Extended): For Sketchy dataset, we use train/test splits proposed by [47] following the Zero-Shot setup. [47] split the dataset into 104 train classes and test 21 classes. Here, the split is done to ensure that there are no common classes in the test set and the ImageNet-1000 dataset. To train our model, we need image and sketch pairs. We randomly select image and sketch from the same training class and pair them. We have 1000 pairs per class in the training set.

TU Berlin Dataset (Extended): TU Berlin is a highly biased dataset. It has some classes with a large number of examples while some with only a few. It is a tough task to learn with the biased data in Zero-shot setting. Therefore, to remove the biases, we equally sample image and sketch pairs from each of the class. While testing, we select the class that has more than 400 samples. As mentioned in the above section, we follow the same pattern to form the image and sketch pairs. Here we have 1500 pairs of image and sketch in each of the class. To have a fair comparison, we randomly select 30 classes for the A_{te} and remaining 220 classes for the A_{tr} as proposed in [37].

4.2. Implementation Details

Our network consists of two Generator and four Discriminator modules. We train the generator G_ϕ to generate attribute(sketch) features and G_θ to generate image features. Whereas, each of the Discriminator distinguishes between samples of images or attributes taken from its corresponding learned joint distributions $p_1(\mathbf{x}, \mathbf{y})$, $p_2(\mathbf{x}, \mathbf{y})$, $p_3(\mathbf{x}, \mathbf{y})$ and $p_4(\mathbf{x}, \mathbf{y})$ and the true joint distribution $q(\mathbf{x}, \mathbf{y})$. We use a series of fully connected layers in all these modules and apply ReLU after each layer except the last layer.

A 300-dimensional noise vector concatenated with 2048 dimensional conditional attribute features $\mathbf{Y}$, is fed into the generator G_θ to learn conditional distribution $p_\theta(\mathbf{x}|\mathbf{y})$. To learn marginal distribution $p_\theta(\mathbf{x})$, we concat 300-dimensional noise vector with 2048-dimensional zero vector and pass it into the generator G_θ. The attribute features $\mathbf{Y}$ is 2048-dimension features of sketches, obtained from ResNet-152 [12]. G_θ passes the input features through a series of 4 FC layers with 1024, 512, 1024, 2048 neurons respectively and outputs 2048-dimensional feature vector $\hat{\mathbf{X}}_\mathbf{1}$ corresponding to the real image $\mathbf{X}$.

Similarly, we concat 300-dimensional noise vector with 2048-dimensional conditional image features, and feed into the generator G_ϕ to learn conditional distribution $p_\phi(\mathbf{y}|\mathbf{x})$. To learn marginal distribution $p_\phi(\mathbf{y})$, we concat 300-dimensional noise vector with 2048-dimensional zero vector and pass it into the generator G_ϕ. G_ϕ passes the input features through a series of 4 FC layers with 1024, 512, 1024, 2048 neurons respectively and outputs 2048-dimensional feature vector $\hat{\mathbf{Y}}_\mathbf{1}$ corresponding to the attribute feature $\mathbf{Y}$.

Discriminator modules D_{ω_2} and D_{ω_3} tries to distinguish between the features of real images $\mathbf{X}$, and features $\hat{\mathbf{X}}_\mathbf{1}$ generated from G_θ while determining joint distribution $p_2(\mathbf{x}, \mathbf{y}) = q(\mathbf{y})p_\theta(\mathbf{x}|\mathbf{y})$ and $p_3(\mathbf{x}, \mathbf{y}) = p_\beta(\mathbf{y})p_\theta(\mathbf{x}|\mathbf{y})$ respectively. Discriminator modules D_{ω_1} and D_{ω_4} tries to distinguish between the attribute features $\mathbf{Y}$, and features $\hat{\mathbf{Y}}_\mathbf{1}$ generated from G_ϕ while determining joint distribution $p_1(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})p_\phi(\mathbf{y}|\mathbf{x})$ and $p_4(\mathbf{x}, \mathbf{y}) = p_\alpha(\mathbf{x})p_\phi(\mathbf{y}|\mathbf{x})$ respectively. All the Discriminator modules takes 2048 dimension feature vectors and passes through a series of 3 FC layers having 1024, 512, and 128 neurons respectively. It outputs the probability of the features being real.

We train our network using Adam Optimizer on $L_{jointGAN}$ loss shown in Equation 13 with learning rate = 0.0001, batch size = 50 and hyper-parameters λ_1=1, λ_2=1 and **z**=300. We choose the hyper-parameters by cross-validation. While training, we first train the generator separately for four epochs and then train the entire network end-to-end for $L_{jointGAN}$ loss. We observe that the validation performance saturates after 30 epochs. We are not using any improved model of the GAN simple GAN [10] model is used without using any gradient penalty.

4.3. Result analysis and Comparison with existing baseline approaches

We compare our proposed model with the existing state-of-the-art of SBIR, ZSL baselines, and recently proposed ZS-SBIR approaches. For a fair comparison, we reproduce the results using the same ResNet-152 features for all the baselines. Table 1 shows the comparison without using any side information.

For a comparison with the existing SBIR techniques we choose Siamese-1 [30], Siamese-2[50], Fine-Grained Triplet [35], Coarse-Grained Triplet[36] as baseline models. We build all the models as per the description in the original paper and train them under our zero-shot setting.

For comparison with ZSL baselines, we choose Direct Regression, ESZSL[32], DAP[18], and SAE[16] as benchmarks and implement these models for sketch-based image retrieval task. We observe that all existing baseline approaches for SBIR and ZSL are not able to perform significantly for unseen class sketches. The main reason of their failure for the unseen classes is that originally these models are trained in a supervised setting and have not used any knowledge transfer learning, so these approaches cannot be leverage for unseen classes.

Recently, CVAE[47], ZSIH[37] and Doodle to search[5] are proposed models for ZS-SBIR. Among these CVAE uses only sketch features as condition attribute to synthesis the image features, whereas remaining others use class description along with sketch features. Class description represents the semantic information about sketches and act as side information to train the model. For fair comparison, we compare our model with CVAE[47]. Since CVAE has conducted the experiments only on Sketchy datasets for VGG-16 features. For CVAE, we reproduce the results for Resnet-152 features on Sketchy and Berlin datasets. In Table 1, we observe that our model outperforms CVAE by 5.0%, 6.2% absolute improvement in precision@200 and mAP@200 respectively in Sketchy dataset and 2.2%, 2.0% absolute improvement in precision@200 and mAP@200 respectively in Berlin dataset.

Figure 4 shows the top 10 retrieval results for unseen class sketches of our proposed model, Y indicates the correct retrieval, and N indicates the false-positive image retrieval. We can observe that our model can generalize for unseen sketch classes. Our proposed model tries to learn the association between sketches and images based on the shape and outline of the sketches instead of the class labels. Our model retrieves some false-positive images for input sketches. We observe that all the false-positive images are very close to the true class images in shape and outline.

5. Ablation Study

We perform ablation over different modules of the proposed model. The proposed model shows a significant improvement in overall baselines. Experimentally we found that our model is robust for the novel classes without using any side information. The detailed ablation is shown in Table 2. The MMD [11] and the CC [40, 53] are the new components added to the jointGAN model. Our ablation shows that the cycle-consistency and MMD is the key component over the jointGAN that boost the performance of the model. We also train our model using class word representation as side information along with sketch features for ablation purpose. We observe a significant performance boost as compared to without using any side information to train the model. Word vector representation of class describes the semantic structure, which helps to transfer the knowledge from seen classes to unseen classes. If we add class representation with a sketch feature, then it guides the generator G_θ to synthesizing semantic preserving and well class discriminative image features.

We perform an extensive analysis of the MMD and CC module. The cycle-consistency preserves the structure of the generated samples in the original space. Therefore from the generated samples, we can reconstruct the attribute/sketch features. This discourages the model from remembering the generated samples hence give the better generalization. The MMD component in-force the model such that the generated samples between the two class have maximum margin, therefore increase the robustness to the retrieval performance.

The joint model with MMD and CC gives a significant boost to image retrieval. Table 2 shows the performance of each of the modules. JointGAN with CC loss shows the 8.7% and 7.7% absolute improvement on the precision@200 and mAP@200 metric and 3.9% and 2.1% absolute improvement on the precision@200 and mAP@200 metric over the Sketchy and Tu-Berlin datasets respectively. If we incrementally add the MMD loss with JointGAN and CC, we observe 1.4% and 0.8% absolute improvement on the precision@200 and mAP@200 metric and 1.6% and 1.3% absolute improvement on the precision. Using side information boost the model performance with a significant margin. Table 2 shows that if we add SI with JointGAN, CC, and MMD, our model gains an absolute improvement of 9.6%, 11.3% in precision@200 and mAP@200 in Sketchy dataset.

6. Conclusion

This paper proposes to use a new generative framework for solving the challenging ZS-SBIR problem which focuses on modeling joint distribution between the image and sketch domains, thus generating high-quality images for

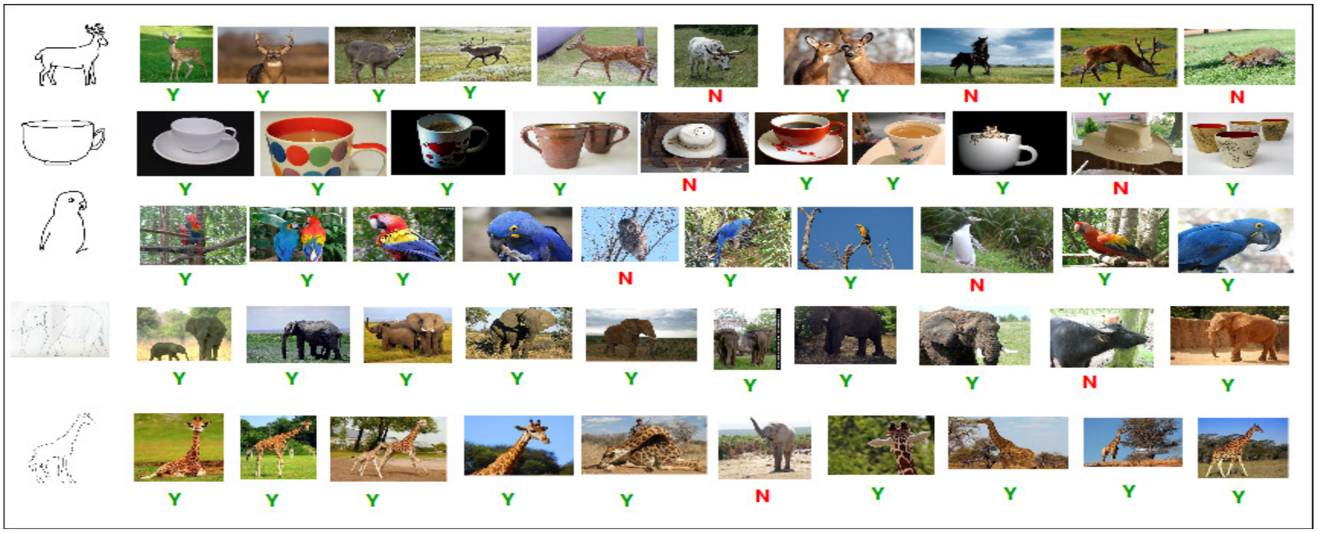

Figure 4. Top 10 Retrieval results of our proposed model. Here we can see that a retrieved object fails when the sketch outline is very close to the image outline. $\mathbf{Y}$ indicates true positive retrieval results and $\mathbf{N}$ indicates false positive retrieval results.

Modules	**Sketchy Dataset**		**TU Berlin Dataset**	
	Precision@200	**mAP@200**	**Precision@200**	**mAP@200**
JointGAN	0.218	0.136	0.149	0.095
JointGAN+CC	0.305	0.213	0.188	0.116
JointGAN+CC+MMD	0.319	0.221	0.204	0.129
JointGAN+CC+MMD+SI	0.415	0.334	0.345	0.264

Table 2. Ablation study for proposed model. Different modules are incrementally added. $\mathbf{CC}$, $\mathbf{MMD}$, $\mathbf{SI}$ corresponds to cycle consistency loss, maximum mean discrepancy loss and Side Information respectively.

an unseen class. The proposed generative model reduces the traditional zero-shot learning problem to the supervised learning problem. We further improve the base model jointGAN with the help of the Maximum Mean Discrepancy loss(MMD). The combined model of the jointGAN with the MMD and CC ensures maximum separability of the generated samples between two classes while preserving the structure. This model surpasses the state-of-the-art system.

References

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

[2] T. Bui, L. S. F. Ribeiro, M. Ponti, and J. P. Collomosse. Generalisation and sharing in triplet convnets for sketch based visual search. *CoRR*, abs/1611.05301, 2016. 1

[3] X. Cao, H. Zhang, S. Liu, X. Guo, and L. Lin. Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor. In *ICCV*, pages 313–320, 2013. 1

[4] Y. Cao, C. Wang, L. Zhang, and L. Zhang. Edgel index for large-scale sketch-based image search. IEEE, 2011. 1

[5] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 7

[6] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31:44–1, 2012. 5

[7] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 34(5):482–498, 2010. 2

[8] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics*, 17(11):1624–1636, 2011. 2

[9] C. Galea and R. A. Farrugia. Forensic face photo-sketch recognition using a deep learning-based architecture. *IEEE Signal Processing Letters*, 24(11):1586–1590, 2017. 1

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 7

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 2, 4, 5, 7

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 6

[13] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1025–1028. IEEE, 2010. 2

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

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

[16] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017. 6, 7

[17] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, June 2018. 2

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 36(3):453–465, 2014. 3, 6, 7

[19] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pages 2862–2871, 2017. 2, 5

[20] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. Adversarial autoencoders. In *ICLR*, 2016. 3

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013. 2

[22] A. Mishra, M. Reddy, A. Mittal, and H. A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. *CVPR-Workshop*, 2017. 2

[23] A. Mishra, V. K. Verma, M. S. K. Reddy, A. Subramaniam, P. Rai, and A. Mittal. A generative approach to zero-shot and few-shot action recognition. *WACV*, pages 372–380, 2018. 2

[24] H. Mohamadi, A. Shahbahrami, and J. Akbari. Image retrieval using the combination of text-based and content-based algorithms. *Journal of AI and Data Mining*, 1(1):27–34, 2013. 1

[25] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2

[26] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2

[27] S. Parui and A. Mittal. Similarity-invariant sketch-based image retrieval in large databases. In *ECCV*, pages 398–414, 2014. 2

[28] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 2, 3

[29] Y. Pu, S. Dai, Z. Gan, W. Wang, G. Wang, Y. Zhang, R. Henao, and L. C. Duke. JointGAN: Multi-domain joint distribution learning with generative adversarial nets. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4151–4160, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 1, 2, 3, 5

[30] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 2460–2464. IEEE, 2016. 1, 2, 6, 7

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

[32] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015. 6, 7

[33] J. M. Saavedra, J. M. Barrios, and S. Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, volume 1, page 7, 2015. 2

[34] J. M. Saavedra and B. Bustos. An improved histogram of edge local orientations for sketch-based image retrieval. In *Joint Pattern Recognition Symposium*, pages 432–441. Springer, 2010. 2

[35] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35:119, 2016. 2, 5, 6, 7

[36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6, 7

[37] Y. Shen, L. Liu, F. Shen, and L. Shao. Zero-shot sketch-image hashing. In *CVPR*, June 2018. 1, 2, 3, 6, 7

[38] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015. 3

[39] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5552–5561. IEEE, 2017. 1

[40] L. C. Tiao, E. V. Bonilla, and F. Ramos. Cycle-consistent adversarial learning as approximate bayesian inference. *arXiv preprint arXiv:1806.01771*, 2018. 4, 7

[41] V. K. Verma and P. Rai. A simple exponential family framework for zero-shot learning. In *ECML-PKDD*, pages 792–808, 2017. 2

[42] Q. Wang and K. Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 124(3):356–383, 2017. 2, 3

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

[44] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pages 4582–4591, 2017. 2

[45] X. Xu, T. Hospedales, and S. Gong. Semantic embedding space for zero-shot action recognition. In *ICIP*, pages 63–67, 2015. 3

[46] X. Xu, T. Hospedales, and S. Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 123(3):309–333, 2017. 2, 3

[47] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal. A zero-shot framework for sketch-based image retrieval. *ECCV*, 2018. 1, 2, 3, 6, 7

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

[49] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016. 1

[50] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Sketch-a-net: A deep neural network that beats humans. *International Journal of Computer Vision*, 122(3):411–425, 2017. 2, 6, 7

[51] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. Sketchnet: Sketch classification with web images. In *CVPR*, pages 1105–1113, 2016. 5

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

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017. 4, 7