Semantic Adversarial Network for Zero-Shot Sketch-Based Image Retrieval

Xinxun Xu\textsuperscript{1}, Hao Wang\textsuperscript{1}, Leida Li\textsuperscript{2}, Cheng Deng\textsuperscript{1}
\textsuperscript{1}School of Electronic Engineering, Xidian University, Xi’an 710071, China
\textsuperscript{2}China University of Mining and Technology
\{xinxun.xu, haowang.xidian.chdeng.xd\}@gmail.com, reader1104@hotmail.com

Abstract

Zero-shot sketch-based image retrieval (ZS-SBIR) is a specific cross-modal retrieval task for retrieving natural images with free-hand sketches under zero-shot scenario. Previous works mostly focus on modeling the correspondence between images and sketches or synthesizing image features with sketch features. However, both of them ignore the large intra-class variance of sketches, thus resulting in unsatisfactory retrieval performance. In this paper, we propose a novel end-to-end semantic adversarial approach for ZS-SBIR. Specifically, we devise a semantic adversarial module to maximize the consistency between learned semantic features and category-level word vectors. Moreover, to preserve the discriminability of synthesized features within each training category, a triplet loss is employed for the generative module. Additionally, the proposed model is trained in an end-to-end strategy to exploit better semantic features suitable for ZS-SBIR. Extensive experiments conducted on two large-scale popular datasets demonstrate that our proposed approach remarkably outperforms state-of-the-art approaches by more than 12\% on Sketchy dataset and about 3\% on TU-Berlin dataset in the retrieval.

1 Introduction

With the explosive growth of image contents on the Internet, image retrieval has played a crucial role in many fields, such as e-commerce, medical diagnosis and remote sensing. Conventional image retrieval requires providing textual descriptions, which are hard to be obtained in some cases. With the rapid development of mobile devices, image retrieval with free-hand sketches, illustrating targeted candidates visually and concisely, has attracted widespread attention and formed the term of Sketch-Based Image Retrieval (SBIR) among the computer vision community. However, it is difficult to guarantee that all categories can be trained in realistic scenarios, which brings about unsatisfactory performance when testing on unseen categories. Hence, Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) is introduced to tackle this practical and challenging problem.

There are two major problems involved in ZS-SBIR: (1) sketches possess large intra-class variance, which significantly increases the complexity of correspondence modeling between natural images and free-hand sketches. For instance, the sketches of a same cat drawn by different people can be obviously distinct. (2) The zero-shot setting requires the model being capable of transferring knowledge from the seen categories to the unseen ones.

To cope with ZS-SBIR comprehensively, both relevant SBIR and zero-shot learning approaches are discussed in this paper. The SBIR approaches [Hu and Collomosse, 2013; Saavedra et al., 2015; Yu et al., 2016; Sangkloy et al., 2016; Chopra et al., 2005; Liu et al., 2017] utilize discriminative representation to model the correspondence between the sketch and the image. However, the performance of these approaches drops significantly under the zero-shot setting as they are essentially designed with the discriminative setup, which encourages class-specific learning.

The zero-shot approaches [Akata et al., 2016; Romera-Paredes and Torr, 2015; Kodirov et al., 2017; Xian et al., 2016] learn an intermediate semantic representation to transfer the knowledge from the seen classes to the unseen ones. However, the primary challenge faced in learning such a mapping is domain shift. Therefore, some researchers proposed to adopt the generative model [Bucher et al., 2017] to tackle this issue. These developments of SBIR and zero-shot learning have facilitated the progress of ZS-SBIR [Kiran Yelamarthi et al., 2018] to a great extent. Unfortunately, they still have a fatal disadvantage for ignoring the large intra-class variance.

\begin{figure}[h]
\centering
\includegraphics[width=0.48\textwidth]{sketch_features.png}
\hspace{0.05\textwidth}
\includegraphics[width=0.48\textwidth]{learned_features.png}
\caption{Visualization both of the sketch features and learned semantic features on Sketchy dataset. We sample 10 classes of sketches from test categories and visualize the distribution of the features. Each color represents a particular class.}
\end{figure}
of sketches, which causes difficulty both in modeling the correspondence and transferring visual knowledge of sketches from the seen categories to the unseen ones. Considering that the word vectors have category-level information, we expect that they are beneficial to diminishing the intra-class variance of sketch features.

In this paper, we propose a novel end-to-end semantic adversarial network to address the large intra-class variance of sketches for ZS-SBIR. Specifically, we elaborate a semantic adversarial module for sketches to learn semantic features with small intra-class variance, which significantly boosts the knowledge transferring of sketches from the seen categories to the unseen ones. As is shown in Figure 1, the intra-class variance of semantic features is smaller than that of sketches by visualizing the distributions of both sketch features and the semantic features with t-SNE [Maaten and Hinton, 2008]. Furthermore, to preserve the discriminability of synthesized features within each training category, a triplet loss is designed for our generative module during the training stage. Besides, an end-to-end learning strategy simplifies the training procedure and also creates the possibility of producing better performance. Extensive experiments on two large-scale popular datasets demonstrate that our proposed approach greatly outperforms state-of-the-art methods by more than 10% on Sketchy dataset and about 3% on TUBerlin dataset.

The main contributions of this work are as follows:

- By taking advantage of category-level semantic information at the training stage, our proposed approach can obtain semantic features of sketches to effectively address its large intra-class variance, which significantly boosts the retrieval performance under zero-shot setting.
- By utilizing the triplet loss on the generative module, our approach can preserve the discriminability of synthesized image features within each category.
- The end-to-end learning strategy of our proposed approach not only simplifies the training procedure, but also brings further improvement in performance.

2 Related Work

2.1 Sketch-Based Image Retrieval

The existing approaches of SBIR can be mainly divided into two categories: the hand-crafted features based methods and the deep learning based ones. The first category includes the gradient field HOG descriptor [Hu and Collomosse, 2013] and the learned key shape (LKS) [Saavedra et al., 2015]. With the development of deep learning, Convolutional Neural Networks (CNN) are widely utilized in computer vision fields. [Yu et al., 2017] first attempted to use CNN for sketch classification. [Chopra et al., 2005] introduced siamese architecture and [Sangkloy et al., 2016] adopted triplet ranking loss for coarse-grained SBIR. [Liu et al., 2017] proposed a semi-heterogeneous deep architecture to learn binary codes of sketches and images. For instance-level SBIR, [Yu et al., 2016] utilized triplet network and evaluated the approach on the shoe and chair datasets.

2.2 Zero-Shot Learning

Due to the expensive cost of data collection and annotation, zero-shot learning has attracted widespread attention in many fields, such as image tagging [Li et al., 2015], visual question answering [Ramakrishnan et al., 2017] and action recognition [Qin et al., 2017]. Existing zero-shot approaches can be divided into two categories: embedding based and generative based approaches. In the first category, ALE [Akata et al., 2016] measured the bilinear compatibility between image embedding space and label embedding space while E2EM [Romera-Paredes and Torr, 2015] and SAE [Kodirov et al., 2017] explicitly regularized the projection between image embedding space and label embedding space. Furthermore, LATEM [Xian et al., 2016] utilized a non-linear component to improve the correspondence modeling between two embedding spaces. As for generative based approaches, [Bucher et al., 2017] proposed a conditional Generative Moment Matching Network (GMMN) to synthesize features of unseen category.

2.3 Zero-Shot Sketch-Based Image Retrieval

To the best of our knowledge, there are only two prior works [Kiran Yelamarthi et al., 2018; Shen et al., 2018] on ZS-SBIR. [Shen et al., 2018] proposed a Zero-shot Sketch-Image Hashing (ZSIH) model consisting of sketches and images binary encoders and a multi-modal learning network to mitigate heterogeneity between modalities. [Kiran Yelamarthi et al., 2018] attempted to synthesize image features of corresponding sketches by utilizing Conditional Variational Autoencoders (CVAE) and then conduct retrieval from image aspect. However, both of them leave the large intra-class variance of sketches out of the consideration.

3 Methodology

3.1 Problem Definition

Assuming \( D_{tr} = \{(ske_i, img_i, wv_i, y_i)\}_{i=1}^{N_{tr}} \) denotes the training set with \( N_{tr} \) samples and \( D_{te} = \{(ske_i, y_i)\}_{i=2}^{N_{te}} \) is the test set with \( N_{te} \) samples, where \( ske_i, img_i, wv_i \) and \( y_i \) are sketch, natural image, word vector and label respectively. Their corresponding label spaces are \( Y_{train} = \{1, 2, 3, ..., s\} \) and \( Y_{test} = \{s + 1, s + 2, ..., s + u\} \), which satisfy the zero-shot setting \( Y_{train} \cap Y_{test} = \emptyset \). The strategy of partitioning the category into \( Y_{train} \) and \( Y_{test} \) is whether it appears in 1000 classes of ImageNet [Deng et al., 2009]. Adopting \( G_\theta \) to model the probability distribution of the image features conditioned on the sketch features (i.e., \( P(img|ske, \theta) \)), the \( G_\theta \) is trained to generate image features using sketch-image pairs from the training classes during the training stage. At testing stage, given \( ske \) belonging to the unseen categories \( Y_{test} \), our proposed approach synthesizes corresponding natural image features to retrieve candidate images from the test image retrieval gallery. The architecture of our proposed model is illustrated in Figure 2, which consists of three components: a semantic adversarial module, an image feature module, a generative module.
3.2 Semantic Adversarial Module

To address the large in-class variance of sketches, we elaborate a semantic adversarial module to learn semantic features with the aid of word vectors that models category-level relationships between different classes. As illustrated in Figure 2, the module has three subnetworks: word embedding network, semantic feature network and discriminative network. Given a training sketch for instance, we first extract sketch features from the conv5 layer of VGG16 [Simonyan and Zisserman, 2014] and obtain category-level word representation from the word vectors model pre-trained on Wikipedia [Mikolov et al., 2013]. Then the goal of this module is to learn semantic features from sketch features in an adversarial fashion [Goodfellow et al., 2014], which means that the semantic features are expected to be as similar as the word vectors by ‘fooling’ a discriminator network \(D_{\theta_D}\). Specifically, the loss function of this module can be formulated as

\[
\mathcal{L}_{adv} = \mathbb{E}_y (\log D_{\theta_D}(W(y))) + \mathbb{E}_{x,y} (\log[1 - D_{\theta_D}(S_{\theta_S}(x^{skc}))]),
\]

where \(x^{skc}, y, W(\cdot), S(\cdot)\) and \(D_{\theta_D}(\cdot)\) denote the sketch, label, word embedding function, semantic embedding function and discriminator function respectively. Besides, the semantic network \(S_{\theta_S}(\cdot)\) and discriminator network \(D_{\theta_D}(\cdot)\) are parameterized by \(\theta_S\) and \(\theta_D\).

3.3 Image Feature Module

As illustrated in Figure 2, we adopt the VGG16 [Simonyan and Zisserman, 2014] pre-trained on ImageNet [Deng et al., 2009] as the image feature extractor and then extract features of fc2 layer. Afterwards, the image features are concatenated with the learned semantic features and subsequently fed into the conditional encoder at the training stage.

3.4 Generative Module

Taking the input of conditional information and random noises, the generative module synthesizes corresponding image features to retrieve natural images by searching the nearest neighbors of them. Our generative module contains one conditional encoder and two decoders. It maps a prior distribution on a hidden latent variable \(P(z)\) to the data distribution \(P(x^{img}, x^{sem})\). The \(P(z|x^{img}, x^{sem})\) is approximated by the variational distribution \(Q(z|x^{img}, x^{sem})\) which is assumed to be Gaussian. Specifically, the variational distribution \(Q(z|x^{img}, x^{sem})\) is approximated via encoder network \(E\) parameterized by \(\theta_E\). The conditional distribution \(P(x^{img}|z, x^{sem})\) is modeled by image decoder network \(D'\) parameterized by \(\theta_{D'}\). Following the notation in [Kingma and Welling, 2013], we can formulate the generative loss as

\[
\mathcal{L}_{KL}(\theta_E, \theta_{D'}; x^{img}, x^{sem}) = \mathcal{KL}(Q_{\theta_E}(z|x^{img}, x^{sem}) || P_{\theta_{D'}}(z|x^{sem})) - \mathbb{E}[\log P_{\theta_{D'}}(x^{img}|z, x^{sem})],
\]

Specifically, \(x^{img}\) and \(x^{sem}\) stand for natural image features and semantic features respectively. Sampling from the standard deviation vector, we add the sample to the mean vector so that we get the sampled latent vector. Then the image decoder takes the features which concatenate latent vector with learned semantic features as input to synthesize corresponding image features at the training stage. Assuming that \(\tilde{x}^{img}\) is the decoder output, the generation of decoder can be formulated as

\[
\tilde{x}^{img} = D'_{\theta_{D'}}(\text{noise}, x^{sem}).
\]

The image reconstruction loss can be formulated as

\[
\mathcal{L}_{recon, img} = \|\tilde{x}^{img} - x^{img}\|_2^2.
\]

Moreover, we connect semantic decoder to reconstruct semantic feature in order to encourage the synthesized image
feature to retain category-level semantic information. Assuming that $\tilde{x}^{\text{sem}}$ is the output of the semantic decoder, the generation of semantic decoder can be formulated as

$$\tilde{x}^{\text{sem}} = R_\theta (\text{noise}, \tilde{x}^{\text{img}}), \quad (6)$$

The semantic reconstruction loss can be formulated as

$$\mathcal{L}_{\text{recon-sem}} = ||\tilde{x}^{\text{sem}} - x^{\text{sem}}||^2_2, \quad (7)$$

where $R$ is the semantic decoder with the parameter of $\theta_R$. It’s noticed that the noise in Eq. (4) and Eq. (6) is the sampled latent vector during the training stage.

### 3.5 Triplet Constraint

To preserve the discriminability of the synthesized features within each training category, inspired by [Karaletsos et al., 2015], we introduce the triplet loss. Concretely, as shown in Figure 3, a triplet consists of an anchor $x^a_i$, a positive $x^p_i$ and a negative sample $x^n_i$. The positive sample shares the same identity with the anchor while the negative sample has a different one. For optimization, a hinge loss is employed to push the negative sample away from the anchor and pull the positive one closer simultaneously. Given a triplet $(x^a_i, x^p_i, x^n_i)$, the objective function can be formulated as

$$\mathcal{L}_{\text{tri}} = \max d(E(x^a_i), E(x^p_i)) - d(E(x^a_i), E(x^n_i)) + \delta, 0), \quad (8)$$

where $d(\cdot, \cdot)$ can be $\ell_2$ distance function, $E(\cdot)$ is latent embedding to obtain $\mu$, and $\delta$ is the margin to be ensured.

### 3.6 Objective and Optimization

The full objective of our proposed model is

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{KL} + \mathcal{L}_{\text{recon-img}} + \mathcal{L}_{\text{recon-sem}} + \mathcal{L}_{\text{tri}}. \quad (9)$$

The whole model is optimized with Adam [Kingma and Ba, 2014] package in PyTorch and the detailed optimization is demonstrated in Algorithm 1.

### 4 Experiments and Discussion

#### 4.1 Datasets and Settings

To perform SBIR under the zero-shot setting, the experiments of this work are taken on two large-scale sketch datasets, i.e., Sketchy [Sangkloy et al., 2016] and TU-Berlin [Eitz et al., 2012]. Details of dataset statistics is in Table 1. To evaluate the performance of the proposed, we follow sketch-based image retrieval evaluation criterion in [Kiran Yelamarthi et al., 2018; Liu et al., 2017], where sketch and image retrieval candidates belonging to same category are regarded as relevant. Specifically, we calculate the average precision (AP) and mean average precision (mAP) for top 200 retrieved candidates.

**Sketchy** is a large-scale sketch dataset, which originally consists of 75,479 sketches and 12,500 images from 125 categories. Then [Liu et al., 2017] extends image gallery by collecting extra 60,502 images from ImageNet [Deng et al., 2009] so that the total number of images in extended dataset is 73,002. Following standard zero-shot setting in [Kiran Yelamarthi et al., 2018; Xian et al., 2018], we partition total 125 categories into 104 train categories as seen classes and 21 test categories as unseen classes according to whether the category appears in 1,000 classes of ImageNet. This partition avoid violating zero-shot assumption when utilizing models that pre-trained on ImageNet.

**TU-Berlin** originally consists of 20,000 unique free-hand sketches evenly distributed over 250 object categories. To perform sketch-based image retrieval, we adopt extended version of TU-Berlin with total 204,070 natural images. Since there is no standard splits setting for ZS-SBIR, we follow the

---

**Algorithm 1** Learning semantic adversarial nets for ZS-SBIR.

**Input:** dataset $D_{tr} = \{(x_i^{sk}, x_i^{im}, x_i^{se}, y_i) | y_i \in Y_{train}\}$, max training iteration $M$ and batch size $N_B$.

**Output:** $\theta_S, \theta_D, \theta_E, \theta_E^N, \theta_R$.

1. Initialize parameters $\theta_S, \theta_D, \theta_E, \theta_E^N, \theta_R$;
2. Create the triple set; $S = \{(T^a_i, T^p_i, T^n_i)\}$, $T_i = (x_i^{sk}, x_i^{im}, x_i^{se}, y_i)$;
3. for $i = 1$ to $M$ do
   4. Calculate adversarial loss $\mathcal{L}_{\text{adv}}$ with Eq. (1);
   5. Forward model to generate $\mu_a, \mu_p, \mu_n$ and calculate the triplet loss $\mathcal{L}_{\text{tri}}$ with Eq. (8);
   6. Forward model to generate $\tilde{x}^{\text{img}}, \tilde{x}^{\text{sem}}$;
   7. Calculate $\mathcal{L}_{KL}, \mathcal{L}_{\text{recon-img}}$ and $\mathcal{L}_{\text{recon-sem}}$ with Eq. (2), Eq. (5) and Eq. (7) respectively;
   8. Update $\theta_R \leftarrow -\frac{\partial}{\partial \theta_R} \mathcal{L}(\mathcal{L})$;
   9. Update $\theta_D^N \leftarrow -\frac{\partial}{\partial \theta_D^N} \mathcal{L}(\mathcal{L})$;
   10. Update $\theta_E \leftarrow -\frac{\partial}{\partial \theta_E} \mathcal{L}(\mathcal{L})$;
   11. Update $\theta_D \leftarrow -\frac{\partial}{\partial \theta_D} \mathcal{L}(\mathcal{L})$;
   12. Update $\theta_S \leftarrow -\frac{\partial}{\partial \theta_S} \mathcal{L}(\mathcal{L})$;
4. end for
14. return solution

---

**Table 1:** Statistics for dataset Sketchy and TU-Berlin.

| Statistics          | Sketchy | TU-Berlin |
|---------------------|---------|-----------|
| Train classes       | 104     | 194       |
| Test classes        | 21      | 56        |
| Train sketches      | 62,785  | 15,520    |
| Test sketches       | 12,694  | 4,480     |
| Train images        | 10,400  | 138,839   |
| Images to be retrieved | 10,453  | 65,231    |
criterion in [Kiran Yelamarthi et al., 2018; Xian et al., 2018] and first manually split total 250 categories into 165 train categories as seen classes and 85 unseen categories as unseen classes. To ensure each unseen class has at least 400 images for retrieval evaluation, we move some categories from unseen classes to seen classes and finally yield 194 seen classes for training and 56 unseen classes for testing. Comparing with Sketchy, TU-Berlin is challenging as there is a large number of unseen classes and large proportion of fine-grained categories.

### 4.2 Implementation Details

We train our model by using Adam on PyTorch with an initial learning rate $\eta = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.99$. The input size of the image is $224 \times 224$. We adopt word embedding model [Mikolov et al., 2013] trained on Wikipedia to extract word vectors whose dimension is 300. To avoid training instability, we alternatively train adversarial module and generative module at optimization of first two epochs. Afterwards, we train the whole model by an end-to-end way. During test stage, we concatenate the learned semantic features with Gaussian noise whose dimension is 1024 and feed this concatenated vector into the image decoder to synthesize the image features which are used to retrieve from test image retrieval gallery.

### 4.3 Comparison with Peer Methods

As ZS-SBIR has rarely been proposed before, which is only two works [Kiran Yelamarthi et al., 2018; Shen et al., 2018] to the best of our knowledge, the quantity of related existing methods is limited. This task can be considered as a combination of SBIR and zero-shot learning. Therefore, we also adopt existing relevant SBIR and zero-shot learning approaches for retrieval performance evaluation. For a fair comparison with our model, we extract the VGG16 $conv_5$ features on the seen sketches and extract the VGG16 $fc2$ features on the seen images as input of the comparison methods. Moreover, all the existing methods were re-implemented in this paper by PyTorch.

**SBIR Methods.** The SBIR baseline computes the pairwise similarity for nearest neighbour search. We build these models [Hadsell et al., 2006; Qi et al., 2016; Sangkloy et al., 2016] according to original papers and train the networks under the zero-shot scenario. Besides, we add an experience that we replace the coarse-grained triplet loss with fine-grained triplet loss in [Sangkloy et al., 2016].

**Zero-Shot Methods.** The direct regression is the baseline of zero-shot approaches, where each feature of the image is learned from the sketch features. Furthermore, we select a set of state-of-the-art zero-shot learning algorithms [Romera-Paredes and Torr, 2015; Kodirov et al., 2017; Kiran Yelamarthi et al., 2018] as benchmarks.

**Results and Analysis.** The performances of all the comparisons under the zero-shot settings on the two datasets are presented in Table 2. We observe that: 1) Under the zero-shot setting, our model significantly outperforms the best comparison by around 12% on Sketchy and 3% on TU-Berlin. 2) To some extent, the SBIR methods based positive-negative samples have the ability to generalize the learned representations to unseen classes while simple calculation of pairwise similarity leading to poor performance. 3) The extra regularization for projection in ESZSL [Romera-Paredes and Torr, 2015] and SAE [Kodirov et al., 2017] promotes them to perform better than direct regression. 4) The proposed network and CVAE [Kiran Yelamarthi et al., 2018] performs better than SBIR methods and projection-based methods illustrates that the generative model performs better than others. 5) The semantic features of low intra-class variance obtained by adversarial learning promote the proposed network to synthesize more effective features than CVAE [Kiran Yelamarthi et al., 2018] in this task.

We show retrieved images for sketch inputs of the unseen classes using the CVAE model and our model in Figure 4. The red border indicates that the retrieved image does not

| Types          | Evaluation Methods                     | Sketchy       | TU-Berlin     |
|----------------|----------------------------------------|---------------|---------------|
|                |                                        | Precision@200 | mAP@200       | Precision@200 | mAP@200       |
| SBIR Methods   |                                        |               |               |               |               |
|                | Pairwise Similarity                    | 0.094         | 0.045         | 0.050         | 0.031         |
|                | Siamese-1 [Hadsell et al., 2006]       | 0.293         | 0.189         | 0.127         | 0.061         |
|                | Siamese-2 [Qi et al., 2016]            | 0.305         | 0.200         | 0.133         | 0.067         |
|                | Coarse-Grained Triplet [Sangkloy et al., 2016] | 0.278     | 0.176         | 0.128         | 0.057         |
|                | Fine-Grained Triplet                   | 0.284         | 0.183         | 0.086         | 0.050         |
| Zero-Shot Methods | Direct Regression                | 0.298         | 0.197         | 0.117         | 0.062         |
|                | ESZSL [Romera-Paredes and Torr, 2015] | 0.305         | 0.202         | 0.131         | 0.072         |
|                | SAE [Kodirov et al., 2017]             | 0.314         | 0.204         | 0.152         | 0.084         |
|                | CVAE [Kiran Yelamarthi et al., 2018]   | 0.333         | 0.225         | 0.165         | 0.104         |
| Ours           | Baseline                               | 0.402         | 0.288         | 0.174         | 0.109         |
|                | End-to-End                             | 0.443         | 0.327         | 0.193         | 0.124         |
|               | End-to-End + Triplet Loss              | **0.453**     | **0.336**     | **0.196**     | **0.129**     |

Table 2: The ZS-SBIR performance compared with existing SBIR and zero-shot approaches re-implemented by ourselves.
belong to the correct class. From these results, we find that the retrieved images closely match the outline of the sketch.

4.4 Ablation Studies
In this section, we conduct ablation studies in terms of the end-to-end structure and triplet loss to further evaluate the effectiveness of our proposed model. The results are exhibited in Table 2. The baseline is separately-trained model which we first train the adversarial semantic module to obtain the semantic features and then take the semantic features as input to the generative module.

End-to-end. As shown in Table 2, the end-to-end network outperforms the baseline by more than 4% on Sketchy and 1.7% on TU-Berlin, which illustrates that training the whole network end-to-end is beneficial to finding better solution.

Triplet loss. Due to the lack of discriminative information in generative model, we attempt to preserve discriminative information by triplet loss. During training, we sample from training set to get anchor, positive and negative sketches. Forwarding the network, we obtain the output of the conditional encoder, which is used to construct triplet loss. The result in Table 2 indicates that triplet loss further improves the performance by adding discriminative information.

4.5 Confusion Matrix
The precision@200 and mean average precision@200 only show the results of instance-level. If we want to dig out which category is insensitive to our model, we can adopt confusion matrix to show whether retrieved images corresponding to their categories. Due to limited space, we do not show the confusion matrix for TU-Berlin as it has more than 50 test classes. As shown in Figure 5, our proposed model is also better than the state-of-art in category-level.

5 Conclusion
We have presented an end-to-end adversarial semantic network to address the problem of zero-shot sketch-based image retrieval more effectively. The adversarial semantic module of the proposed network incorporates semantic feature network with the category-level information of word vectors, and carries out adversarial learning to maximize the semantic relevance and feature distribution consistency between semantic features and word vectors. Therefore, the adversarial module diminishes the intra-class variance of the input features of the generative module, which achieves better performance. Moreover, the triplet loss can preserve the discriminability of synthesized features within each training category. Last but not least, compared with training the network separately, training the whole network in an end-to-end fashion is beneficial to finding better solution, which endows our network with zero-shot generalization ability. Experiments on two large datasets verified our proposed model significantly outperforms existing methods in ZS-SBIR task.
References

[Akata et al., 2016] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1425–1438, 2016.

[Bucher et al., 2017] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. In *ICCV Workshop on TASK-CV: Transferring and Adapting Source Knowledge in Computer Vision*, pages 2666–2673, 2017.

[Chopra et al., 2005] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546, 2005.

[Deng et al., 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Eitz et al., 2012] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44–1, 2012.

[Goodfellow et al., 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

[Hadsell et al., 2006] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742, 2006.

[Hu and Collomosse, 2013] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Comput. Vis. Image Understand.*, 117(7):790–806, 2013.

[Karaletos et al., 2015] Theofanis Karaletos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. arXiv:1506.05011, 2015.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv:1312.6114, 2013.

[Kiran Yelamarthi et al., 2018] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, pages 300–317, 2018.

[Kodirov et al., 2017] Elyor Kodirov, Tao Xiang, and Shao-gang Gong. Semantic autoencoder for zero-shot learning. arXiv:1704.08345, 2017.

[Li et al., 2015] Xirong Li, Shuai Liao, Weiyu Lan, Xiaoyong Du, and Gang Yang. Zero-shot image tagging by hierarchical semantic embedding. In *ACM SIGIR*, pages 879–882, 2015.

[Liu et al., 2017] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pages 2862–2871, 2017.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Mikolov et al., 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

[Qi et al., 2016] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, pages 2460–2464, 2016.

[Qin et al., 2017] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *CVPR*, pages 2833–2842, 2017.

[Ramakrishnan et al., 2017] Santhosh K Ramakrishnan, Ambar Pal, Gaurav Sharma, and Anurag Mittal. An empirical evaluation of visual question answering for novel objects. In *CVPR*, pages 4392–4401, 2017.

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

[Saavedra et al., 2015] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (LKS). In *BMVC*, volume 1, pages 1–11, 2015.

[Sangkloy et al., 2016] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 35(4):119–130, 2016.

[Shen et al., 2018] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, pages 3598–3607, 2018.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[Xian et al., 2016] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[Xian et al., 2018] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018.

[Yu et al., 2016] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, pages 799–807, 2016.

[Yu et al., 2017] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *Int. J. Comput. Vis.*, 122(3):411–425, 2017.