Bi-Directional Domain Translation for Zero-Shot Sketch-Based Image Retrieval

Anonymous CVPR submission

Paper ID ****

Abstract

*Sketch-Based Image Retrieval (SBIR) is a cross-modal retrieval task, which uses free-hand sketches to retrieve from natural image gallery with the **same category**. However, SBIR requires all categories can be seen during training, which is not guaranteed. So we investigate a more challenging task, Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR), in which testing categories do not appear at training stage. Traditional SBIR method are prone to be category-based retrieval and cannot generalize well from seen categories to unseen ones. In contrast, we disentangle image feature into structure feature and appearance feature to facilitate structure-based retrieval. To realize the function of feature disentanglement and take full advantage of disentangled information, we propose Bi-directional Domain Translation for zero-shot sketch-based image retrieval (BDT) framework, in which the image domain and sketch domain can be translated to each other through disentangled structure and appearance feature. Moreover, we also perform retrieval in both structure feature space and image feature space. Extensive experiments demonstrate that our proposed approach remarkably outperforms state-of-the-art approaches by about 8% on Sketchy dataset and about 5% on TU-Berlin dataset in the retrieval.*

1. Introduction

In recent years, with the rapid growth of multimedia data on the internet, the role of image retrieval has become more and more important in many fields, such as remote sensing and e-commerce. Since the sketch can be easily drawn and express shape and pose details about the target images, sketch-based image retrieval (SBIR), which use a sketch to retrieve images with the same category, has become widely accepted among users. Therefore, SBIR has also attracted widespread attention in the research community [9, 3, 14, 15, 2, 21, 68, 20, 1, 46, 24, 58, 50, 34, 63, 48, 51, 40]. In the conventional setting, it is assumed that the images and sketches in training and testing set share the same categories. However, in real-world applications, the categories of test sketches/images may be out of the scope of training categories.

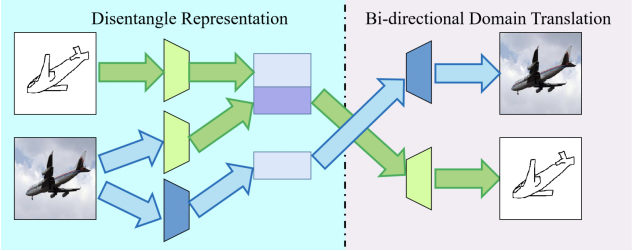

Figure 1: The overall structure of our model

In this paper, we investigate a more challenge task called zero-shot sketch-based image retrieval (ZS-SBIR), which assure that the test categories do not appear at training stage. It has been shown that the performance of existing methods will decline dramatically at ZS-SBIR setting [62], which probably because existing methods are prone to learn category-based retrieval. Specifically, since the evaluation methodology is category-based, traditional SBIR methods may take a shortcut by correlating sketches/images with their category labels and retrieving the images from the same category as the query sketch [62], which is very effective when testing data share the same categories as training data. However, they often fail when the testing categories have never been seen at training stage.

To generalize well from seen categories to unseen categories, a model should learn to align the structure information of sketches with the corresponding structure information of images, which is referred to as structure-based retrieval. Existing methods for ZS-SBIR can be categorized into three groups, (1) some works use generative model based on aligned data pairs, where image and sketch have same shape and pose, to reduce the gap between seen and unseen categories [62]; (2) some works use semantic information to reduce the intra-class variance in sketch to stabilize the training process [60, 59, 12, 52]; (3) some works fine-tune the pre-trained model in ZS-SBIR task, with semantic-aware knowledge preservation to pre-

vent catastrophic forgetting [37]. However, the aligned data pairs and semantic information is not always available. Moreover, all the above methods did not achieve the goal of structure-based retrieval. Apart from the methods designed for ZS-SBIR, some prior works tend to generate structure information based on sketch tokens, which are obtained by directly extracting the outlines of image [36, 58, 63]. However, the sketch tokens obtained in this way are not very reliable due to lots of noisy and redundant information, which limits the performance of these methods. Different from them, we disentangle structure feature from image feature to facilitate structure-based retrieval.

As is known to all, sketch is only an outline of image, so we can disentangle the image feature into structure-features and appearance-features, where the structure-features corresponding to the structure information of outlines and the appearance-feature corresponding to additional detailed information, like background and color. To realize the function of feature disentanglement and take full advantage of disentangled information, we propose Bi-directional Domain Translation for zero-shot sketch-based image retrieval (BDT) framework, in which sketch and image are treated as two domains. As shown in Figure 2, we first use a pre-trained model (e.g., VGG) to extract features from sketches and images, which are referred to as sketch features and image features respectively. Then, two independent encoders and and an orthogonal loss are adopted to disentangle the image features into structure features and appearance features. Sketch feature is also projected to the same structure feature space. Then, bi-directional domain translation is performed through these features. For image-to-sketch translation, we use the image structure features to directly generate the sketch feature. For sketch-to-image translation, we adopt an image generator to reconstruct image feature based on both sketch feature and appearance feature, where the latter is used to compensate the uncertainty during generating the image features from sketch features. Due to the need of stochastic sampling, a variational estimator is adopted to image appearance feature before it combine with sketch structure feature. Moreover, we also perform retrieval in both structure feature space and image feature space. The overall model and the new retrieval strategy is verified by the quantitative experimental results.

Our main contributions of this paper are summarized as follow:

- To our best knowledge, we are the first to disentangle image feature into structure feature and appearance feature, to facilitate structure-based retrieval.
- We propose a bi-directional domain translation framework for zero-shot sketch-based image retrieval task.
- Experimental results on two popular large-scale datasets show that our approach significantly outperforms state-of-the-art methods.

2. Related Work

2.1. SBIR and ZS-SBIR

The main goal of sketch-based image retrieval (SBIR) is to build a bridge between image domain and sketch domain. Basically, the methods to solve this problem can be categorized into hand-crafted features based and deep-learned features based. Before deep learning was introduced to this task, hand-crafted based methods mostly work by extracting the edge map from natural image and then matching them with sketch using different Bag-of-Words model on specifically designed feature [50, 20, 15, 21, 14]. In recent year, deep learned based methods become popular in this area. To reduce the gap between image domain and sketch domain, variant of siamese networks [48, 51, 56] and ranking losses [8, 51] are adopted to this task. Besides, semantics information and adversarial loss are also introduced to preserve the domain invariant information [4].

The zero-shot sketch-based image retrieval is proposed by [52] and then followed by [62, 60, 59, 37, 12]. To reduce the intra-class variance in sketch and stabilize the training process, semantic information is leveraged in [59, 52, 60, 12]. To reduce the gap between seen and unseen categories, generative model along with aligned data pairs is proposed in [62]. To adapt the pre-trained model to ZS-SBIR without forgetting the knowledge of ImageNet [10], semantic-aware knowledge preservation mechanism is used in [37] to prevent the catastrophic forgetting.

2.2. Disentangled Representation

The disentangled representation is to divide the latent representation into multiple units, with each unit corresponding to one latent factor (e.g., the pose feature can be disentangled from face recognition images [57]). Each unit is only affected by its corresponding latent factor, but not influenced by other latent factors.

Disentangled representation is more generalizable and semantically meaningful, and thus useful for a variety of tasks. These methods can be categorized into unsupervised learning and supervised learning. For the unsupervised disentanglement, it includes InfoGAN [6], MTAN [38], β-VAE [19], JointVAE [11], FactorVAE [26], InfoVAE [66] and TCVAE [5]. In this manner, the most common approach first define the prior over the latent variables with a fully-factorized Gaussian distribution and use the prior to encourage the feature disentanglement in different dimensions of the latent representation. For the second part, Kingma *et al.* [29] first use disentangled representation to enhance the semi-supervised learning. Zheng *et al.* [67] propose DG-Net to integrate discriminative and generative learning us-

ing disentangled representation. Besides, the effectiveness of supervised disentanglement in different applications, like person re-id [67], face recognition [35, 39, 53, 57] and image generation [41, 61, 43, 25] also helps to attracted great attention.

2.3. Domain Translation

Many domain translation approaches, like Pix2Pix [23], CycleGAN [69], BiCycleGAN [70], StarGAN [7], DiscoGAN [27] have been proposed, which can translate figures between different domains (e.g., sketch domain and image domain). In this subsection, we mainly discuss the domain translation methods based on disentangled representation. They disentangle latent representation into domain-specific representation and domain-invariant representation. In our problem, structure feature can be treated as domain-invariant representation and appearance feature can be treated as domain-specific feature. DRIT [32] and DRIT++ [33] use the disentangled representation with VAE to enrich the diversity during inference. MUNIT [22] and CDD [17] combine the disentanglement objective into image-to-image translation between two domains and the disentangled representation also help the model to capture domain-specific and domain-invariant information while generating the images, which is similar to our method.

In this paper, we introduce disentangled representation into ZS-SBIR tasks along with a bi-directional domain translation model to take full advantage of disentangled information.

3. Methodology

In this section, we introduce our proposed Bi-directional Domain Translation for zero-shot sketch-based image retrieval (BDT) framework. In Sec 3.1 we state the problem of ZS-SBIR and define some notations in this paper. In Sec 3.2, we elaborate the disentangled representation and the bi-directional domain translation modules in detail. In Sec 3.3 we discuss the strategy during training and retrieving.

3.1. Problem Definition

In this paper, we focus on solving the problem of hand-free sketch-based image retrieval under zero-shot setting, where only the sketches and images from seen categories are used at training stage. At the testing stage, our proposed framework is expected to use the sketches to retrieve the images, the categories of which have never appeared during training.

We first provide a definition of the SBIR in zero-shot setting. Given a sketch dataset $S_{sk} = \{(x_i^{sk}, y_i)|y_i \in \mathcal{Y}\}$ and an image dataset $S_{im} = \{x_j^{im}, y_j|y_j \in \mathcal{Y}\}$, where (x_i^{sk}, y_i) and (x_j^{im}, y_j) are corresponding to the image, sketch and their corresponding category labels. Following the zero-shot setting in [62, 59], we split all categories $\mathcal{Y}$ into $\mathcal{Y}_{train}$ and $\mathcal{Y}_{te}$, where no overlap exists between two label set, i.e. $\mathcal{Y}_{tr} \cap \mathcal{Y}_{te} = \emptyset$. Considering that we use ImageNet pre-trained model to extract images/sketches feature. To avoid the test category shown at training stage, we also require the testing category not in ImageNet. Based on the partition of label set $\mathcal{Y}$, we can split sketch (resp., image) dataset into S_{sk}^{tr} and S_{sk}^{te} (resp., S_{im}^{tr} and S_{im}^{te}) At training stage, our model can only process the data in S_{sk}^{tr} and S_{im}^{tr}. During test, given a sketch s^{sk} from S_{sk}^{te}, our model need to retrieve images belonging to the same category as s^{sk} from test images gallery S_{im}^{te}.

The overall framework of our method is illustrated in Figure 2. Let $f_{im} \in \mathbb{R}^n$ be the image feature, $f_{sk} \in \mathbb{R}^n$ be the sketch feature, $f_{im}^{st} \in \mathbb{R}^m$ be the image structure feature, $f_{im}^{ap} \in \mathbb{R}^m$ be the appearance feature, z_{im}^{ap} be the variational appearance feature and $f_{sk}^{st} \in \mathbb{R}^m$ be the sketch structure feature. Our bi-directional domain translation model builds several feature maps among these features. Before the translation, we first adopt feature extractor to map the image/sketch to their corresponding feature. Two image encoder is applied to disentangle the image feature into structure and appearance features. Meanwhile, the sketch feature is also projected into structure space. For the image-to-sketch translation, the image structure feature is directly translated into the sketch feature. Whereas, for the sketch-to-image translation, we combine the sketch structure feature and the image appearance feature together to get the image feature.

3.2. BDT

3.2.1 Feature Extractor

Since sketches are highly abstract and possess intrinsic visual sparsity compared with natural images, it is hard to extra to get more information from the pre-trained model. To alleviate this problem without adding more parameters, we use the feature fusion network in *Wang et al.* [59] and concatenate the features extracted from multiple layers of the pre-trained model to get the feature of images and sketches.

In detail, we first use a pre-trained backbone model, *viz* VGG-16 in this paper, on ImageNet [10] to process each sketch and image. Suppose $F_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ is the output of the i-th convolution layer and $f_{fc} \in \mathbb{R}^N$ is the feature of the last fully connected layer, the extracted feature f can be obtained by concatenating f^{fc} and global average pooling (GAP) of F^i:

$$f = [f_{fc}, GAP(F_5), GAP(F_4), GAP(F_3)], \quad (1)$$

3.2.2 Disentangled Representation

To achieve the goal of structure-based retrieval, we tend to disentangle structure information from image feature. We

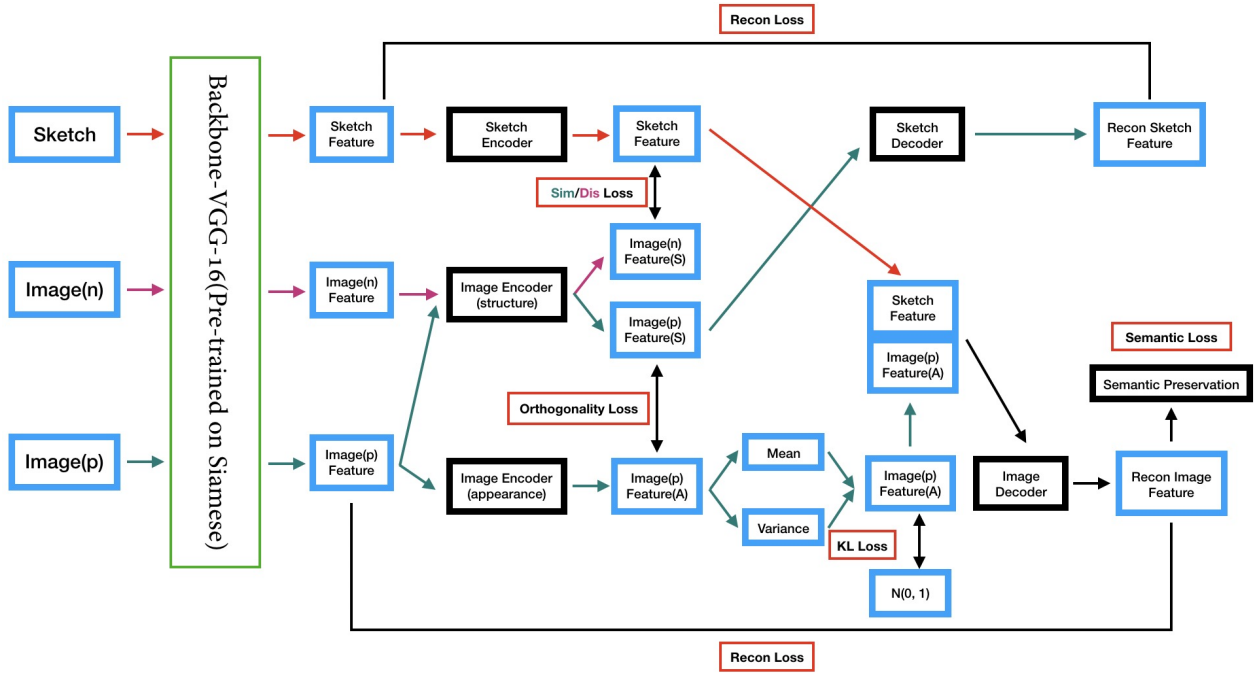

Figure 2: Model Structure

adopt two image encoder to disentangle the image feature into image structure feature and image appearance feature. Besides, to compare the image structure information with sketch in the same space, a sketch encoder is also adopt to get the sketch structure feature.

$$f_{im}^{ap} = E_{im}^{ap}(f_{im}); f_{im}^{st} = E_{im}^{st}(f_{im}); f_{sk}^{st} = E_{sk}^{st}(f_{sk}). \quad (2)$$

At each training step, apart from sampling a positive sketch-image pair (f_{sk},f_{im^+}) from the same category, we also sample another image as negative sample (f_{im^-}), which belongs to other categories. In structure feature space shared by sketch and image, we expect to pull a sketch close to the images of the same category and push a sketch apart from image of a different category. To achieve this, we design a ranking loss in structure feature space to map the disentangled image structure features to the same space as sketch structure feature. Here we use L_2 distance as our metric.

$$\mathcal{L}_{rk} = ||f_{sk} - f_{im^+}^{ap}||_2 + \max(0, m - ||f_{sk} - f_{im^-}^{ap}||_2). \quad (3)$$

Since the negative image sample is not used in the following modules, we use f_{im} to represent f_{im^+} in the following part of this paper.

To make sure the image features are disentangled in the structure feature space and appearance feature space. We impose a orthogonality constrain between image structure feature and image appearance to force them orthogonal under cosine metric.

$$\mathcal{L}_{or} = cos(f_{im}^{ap}, f_{im}^{sk}) = \frac{f_{im}^{ap} \bullet f_{im}^{st}}{||f_{im}^{ap}||_2 ||f_{im}^{ap}||_2} \quad (4)$$

Note that the f_{img}^{ap} and f_{img}^{sk} are the output of ReLU activation so the $cos(f_{im}^{ap}, f_{im}^{sk})$ will always larger than 0.

3.2.3 Bi-directional Domain Translation

To further help the model learn disentangled representation and fully utilize the disentangled image features, we design a the bi-directional domain translation model, which translate the image structure feature to sketch feature and translate the sketch structure feature to image feature at the same time. To reduce the ambiguous aspects of the image feature which are not present in the sketch feature, we adopt a variational estimator to approximate the image appearance feature and integrate it with sketch structure feature to get the image feature.

For the image-to-sketch translation, we directly use the image structure feature to reconstruct sketch feature,

$$\hat{f}_{sk} = G_{sk}(f_{im}^{st}). \quad (5)$$

To measure the quality of the translated results, we adopt a reconstruction loss here. To make the translated sketch feature distribution similar to the original sketch feature distribution, we also add an adversarial loss to constrain the

translation results.

$$\mathcal{L}_{tl}^{sk} = \log(D_{sk}(\hat{f}_{sk})) + ||f_{sk} - \hat{f}_{sk}||_2. \quad (6)$$

For the sketch-to-image translation, we try to use the sketch structure feature to reconstruct image feature as well. Whereas, considering that the images contain much more appearance information than the sketches, it is hard to compensate the appearance uncertainty from the generator only. Therefore, image appearance feature should be integrate to sketch structure feature to reduce the ambiguous. To facilitate the stochastic sampling, a variational estimator V_{img}^{ap} is add to approximate the variational distribution $Q(z_{img}^{ap}|f_{img})$, which is assumed to be Gaussian distribution with $\mu = 0$ and $\sigma = 1$. The conditional distribution $P(f_{img}|z_{img}^{ap}, f_{ske}^{st})$ is modeled by a conditional generator G_{img},

$$\mu_{im}^{ap}, \sigma_{im}^{ap} = V_{im}^{ap}(f_{im}^{ap}), \quad (7)$$

$$z_{im}^{ap} = \mu_{im}^{ap} + \epsilon * \sigma_{im}^{ap} \quad (8)$$

$$\hat{f}_{im} = G_{im}([z_{im}^{ap}, f_{sk}^{st}]) \quad (9)$$

where we use reparameterization trick [30] to sample z_{im}^{ap}, in which ϵ is sampled from the $\mathcal{N}(0,1)$. To enforce the variational distribution $Q(z_{img}^{ap}|f_{img})$ to be close to prior distribution $\mathcal{N}(0,1)$, a Kullback-Leibler Divergence (KLD) is applied to between them. The loss function can be expressed as:

$$\mathcal{L}_{KL} = -D_{KL}(\mathcal{N}(\mu_{im}^{ap}, \sigma_{im}^{ap})||\mathcal{N}(0,1)). \quad (10)$$

Similar to image-to-sketch translation, the loss for sketch-to-image translation can be expressed as:

$$\mathcal{L}_{tl}^{im} = \log(D_{im}(\hat{f}_{im})) + ||f_{im} - \hat{f}_{im}||_2, \quad (11)$$

Besides, the discriminators are trained to distinguish the generated images/sketches from the real ones. So the loss function for discriminators are:

$$L_D^{im} = \log(1 - D_{im}(\hat{f}_{im})) + \log(D_{im}(f_{im})) \quad (12)$$

$$L_D^{sk} = \log(1 - D_{sk}(\hat{f}_{sk})) + \log(D_{sk}(f_{sk})) \quad (13)$$

3.3. Training and Retrieval

The full objective function can be divided into two categories, the generation loss and the discrimination loss, which can be expressed as:

$$\mathcal{L}_G = \lambda_1\mathcal{L}_{or} + \lambda_2\mathcal{L}_{rk} + \lambda_3\mathcal{L}_{KL} + \lambda_4\mathcal{L}_{tl}^{im} + \lambda_5\mathcal{L}_{tl}^{sk}, \quad (14)$$

$$\mathcal{L}_D = \lambda_6 L_D^{im} + \lambda_7 L_D^{sk}, \quad (15)$$

λ are hyper-parameters for balancing the overall performance. Our model consists of generators and discriminators, in order to stabilize the training process, we follow the training strategy in GAN [18] to update them alternatingly with N_D and N_G times respectively.

At test stage, we divided our process into two parts.

1. The image decoder is used to synthesize N image features vectors $\hat{f}_{im}^{i}$ conditioned on sketch structure features and latent vector sampled from $\mathcal{N}(0,1)$. Then the average of the synthesized features is obtained to represent the final synthesized image features $\hat{f}_{im}^{re}$,

$$\hat{f}_{im}^{re} = \frac{1}{n}\sum_{i=0}^{N} G_{im}([f_{sk}, z_i]), \quad (16)$$

   where z_i is sampled from $\mathcal{N}(0,1)$

2. the image structure encoder and the sketch encoder will map the images feature and sketches feature into a structure feature space as f_{im}^{st} and f_{sk}^{st}.

While retrieving the image, we calculate the cosine distance in structure feature space as 1-$cos(f_{im}^{st}, f_{sk}^{st})$ and that in image feature space as 1-$cos(\hat{f}_{im}^{re}, f_{im})$. Then, we use the weighted average of both as the final distance for retrieval.

$$\mathcal{D}_{fusion} = \omega(1 - cos(\hat{f}_{im}^{re}, f_{im})) + (1-\omega)(1 - cos(f_{im}^{st}, f_{sk}^{st})), \quad (17)$$

where ω is the hyper-parameter for balancing between these two feature space.

4. Experiment

4.1. Experiment Setup

4.1.1 Dataset

We evaluated BCD-SBIR on two large-scale sketch-image datasets: TU-Berlin [13] and Sketchy [51] with extended images obtained from [36].

Sketchy (Extended) [51] originally comprised 75,479 sketches and 12,500 images from 125 categories, where the image and sketch are aligned data pairs. Liu *et al.* [36] extended the image retrieval gallery by collecting extra 60,502 images, so that the total number of images in extended Sketchy is 73,002. Following the standard zero-shot setting in [62], we partition the total 125 categories into 104 training categories as seen categories and 21 test categories as unseen categories according to whether the category appears in the 1,000 classes of ImageNet-1k [10], which avoids violating the zero-shot assumption when utilizing models that are pre-trained on ImageNet-1k. At training stage, there are two methods to use the training data, aligned data pair and unaligned data.

TU-Berlin (Extended) [13] contains 250 categories with a total of 20, 000 sketches extended by [36] with natural images corresponding to the sketch classes with a total size of 204,489. Follow the same split criteria as Sketchy, we first

[1] We do not compare the ZSIH [52] in this table

	Method	Sketchy Ext. (aligned)		Sketchy Ext. (Unaligned)		TU-Berlin Ext.	
		P@200(%)	mAP@200(%)	P@200(%)	mAP@200(%)	P@200(%)	mAP@200(%)
SBIR	Cosine	9.0	5.1	-	-	4.6	2.0
	3D shape [58]	6.1	1.0	7.0	1.8	3.6	0.5
	SaN [64]	15.3	5.8	18.9	8.5	10.1	4.2
	Siamese [62]	19.3	9.8	22.1	13.9	8.1	3.7
ZSL	ESZSL [49]	18.6	10.0	18.1	10.0	6.9	3.1
	SAE [31]	24.4	14.4	27.1	17.5	11.6	5.5
	CMT [55]	9.1	2.2	8.6	1.9	3.8	0.5
	SSE [65]	6.9	2.3	7.3	3.3	4.1	1.2
	DeViSE [16]	9.3	1.9	9.0	1.7	3.1	0.3
ZS-SBIR	CVAE [62]	33.4	22.6	31.2	19.9	10.2	4.9
	PCYC [11]	28.0	17.7	30.0	19.4	12.4	5.7
	Xu *et al.* [60]						
	BDT-St	36.1	25.5	36.9	25.8	15.2	7.9
	BDT-Im	37.2	26.8	35.1	24.9	14.7	7.1
	BDT	41.2	29.9	39.7	28.1	17.6	10.2

Table 1: ZS-SBIR performance comparison of BCD and existing methods. - means the results for the result for the two setting in Sketchy are the same. For PYCY[11], we remove the semantics information when training the model. For 3D shape [58], SaN [64], ESZSL [49], SAE [31], CMT [55], SSE [65], DeViSE [16] and Xu *et al.* [60], we replace the semantic information with the average of image features which share the same categories[1].

split the TU-Berlin into 165 training categories and 85 testing categories. As the Shen *et al.* [52] suggest, we re-select testing categories with more than 400 images from the 85 categories. In the end, the training and testing categories of TU-Berlin are 186 and 64 respectively. Compared with Sketchy dataset, TU-Berlin is much more challenging as it contains more unseen categories and relatively less training sketches.

4.1.2 Implementation Details

We implement our model and all the other baseline models using the popular deep learning toolbox PyTorch [47], which are all trained on one GTX 1080Ti GPU. We use a VGG-16 (pre-trained on ImageNet dataset) to extract the image and sketch features. As Sec. 3.2 has mentioned, we concatenate the output of middle layers after the global average pooling (GAP) and get the 5568-D feature for each image and sketch. For each encoder, there are two fully-connected layers with Batch Normalization and ReLU activation, and one Dropout layer between them. For the variational estimator, there are two fully-connected layers working parallel to calculation the mean and the variance of the approximated z_{im}^{ap} respectively. For each decoder, there are two fully-connected layers with ReLU activation. And for the discriminators, there are two fully-connected layers with Batch Normalization and LeakyReLU as activation. The dimension size for f_{sk}^{st}, f_{im}^{st}, f_{im}^{ap}, z_{im}^{ap} are all 1024-D.

We use an Adam [28] optimizer with learning rate of 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$ for optimization across our model except for the discriminators and use SGD optimizer with learning rate of 1×10^{-2}, momentum=0.9 for optimization across the two discriminators. The batch size for Sketchy, batch size for TU-Berlin and maximum number of training epochs are 128, 64, 30 respectively. The training iteration of generator (N_G) and discriminator (N_D) are 100 and 50 accordingly. Note that in Sketchy dataset, we perform experiments on both unaligned data and aligned data, whereas, we only use unaligned data in TU-Berlin for the lack of aligned data.

4.2. Comparison with Existing Methods

We compare our model with twelve prior works, which can be divided into three categories, the sketch-based image retrieval (SBIR) methods, the zero-shot learning (ZSL) methods and the zero-shot sketch-based image retrieval (ZS-SBIR) methods. The SBIR methods that we evaluate are Siamese [62], SaN [64], 3D shape [58] and DSH [36]. A cosine baseline is also added, which compute the 4096-D VGG-16 [54] feature vector pre-trained on ImageNet-1k for nearest neighbor search. The ZSL methods that we evaluate are ESZSL [49], SAE [31], CMT [55], SSE [65] and DeViSE [16]. The ZS-SBIR methods that we evaluate are CVAE [62], PCYC [12] and Xu *et al.* [60]. For fair comparison, we replace the backbone of all the previous model to VGG-16, except for the SaN, which propose a new backbone to extract information from sketch and image. All the backbones are pre-trained on ImageNet-1k. Since we do not use additional semantic information obtained from

	Sketchy Ext. (aligned)		Sketchy Ext. (unaligned)	
	P@200(%)	mAP@200(%)	P@200(%)	mAP@200(%)
-L_{rk}	35.1	23.2	31.7	20.3
-L_{or}	40.3	29.1	38.4	26.9
-L^{im}_{recog}	31.7	19.8	32.1	20.5
-L^{sk}_{recog}	39.9	28.3	38.3	27.8
-L^{im}_{adv}	40.0	28.3	35.5	24.9
-L^{sk}_{adv}	40.7	29.6	39.3	27.9
moved L_{or}	39.1	27.4	37.9	26.6
sk2im	37.2	26.0	36.5	25.9

Table 2: Ablation Study on Sketchy Ext. moved $\mathcal{L}_{or}$ means that we move the orthogonality loss from $(f^{ap}_{im}, f^{st}_{im})$ to $(z^{ap}_{im}, f^{st}_{im})$; sk2im means we directly translate the image from the sketch structure feature and ignore the variational estimator.

large textual corpus (e.g., word vector [44] and WordNet [45]) in our model, we also remove the semantics information or replace the semantics information to the average of image features, which share the same category, for all the baselines. Besides, we do not compare with the methods that need to fine-tune the pre-trained backbone during training the retrieval model, like SAKE [37] and EMS [40]. We use mean average precision and precision considering top 200 (mAP@200 and P@200) retrievals for evaluation and comparison.

Table 1 shows the results of our proposed BDT and all the comparison methods. From the results, we can find that most of the SBIR and ZSL methods perform worse than the ZS-SBIR methods. Compare with the results of Cosine, the 3D shape [58], CMT [55], SSE [65] and the DeViSE [16] perform ever worse, which indicates these model heavily overfit in seen categories. In Sketchy Ext. dataset, we find the model train on the unaligned data usually perform better than the aligned data, which mainly because that the number of unaligned data are five times more than that of aligned data and large training data can help the model better generalize from seen to unseen categories. Whereas, CVAE shows the opposite tendency, as in this case, the aligned data with pose similarity can help the model learn structure-based retrieval when reconstruct the image from the sketch. For the TU-Berlin Ext. dataset, the number of the unseen categories under our split are two times more than that in prior works [37, 12] and our split also guarantee that there is not information leak from unseen categories, so that the results in our paper is much more lower than that in prior works [37, 12].

Our proposed BDT has excelled the state-of-the-art methods by 7.8% P@200 in Sketchy Ext. (aligned) dataset, 8.5% P@200 in Sketchy Ext. (unaligned) dataset and 5.2% P@200 in TU-Berlin Ext. dataset. To better realize the advantage of our proposed method, we also list the retrieval results from only the image feature space and the structure space as **BDT-Im** and **BDT-St**. Compare the results between **BDT-Im** and CVAE and the results between **BDT-St** and Siamese, the disentangled representation is indeed help the model to generalize from seen to unseen categories. Besides, the results among **BDT-Im**, **BDT-St** and **BDT** show that retrieval from both image feature space and structure feature space can boosting the performance by a large margin, which also indicates that retrieval results of these two space can compensate each other.

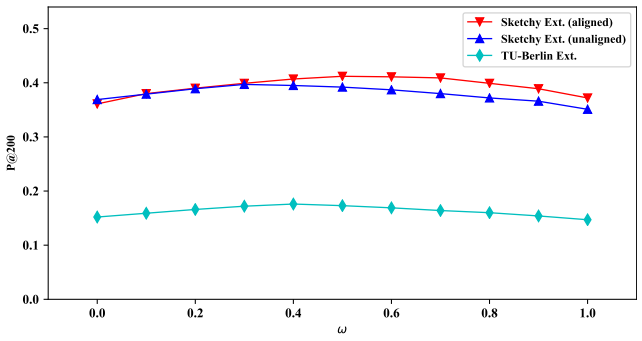

Figure 3: The performance variety along with the ω

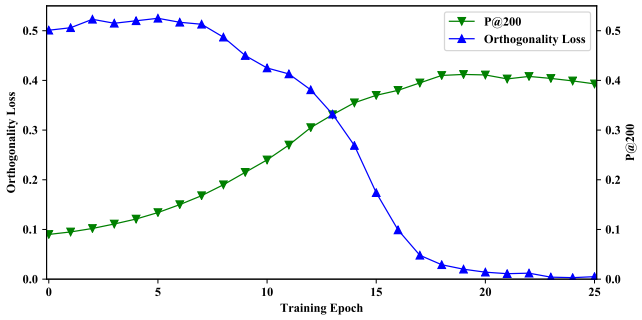

Figure 4: The performance and orthogonality variety along with the training epoch

4.3. Ablation Study

We analyze the effect of different loss functions, some alternations while designing the model and the effect of ω on Sketchy Ext dataset.

Loss Analysis In Table 2, we can find the effect of different losses. As expected, the ranking loss and the image reconstruction loss are the most important constrains in our model, which is because these two losses mainly control the image-sketch distance in their corresponding space. Besides, the image reconstruction loss has bigger effect on aligned data than that on unaligned data, which indicated the reconstruction loss is sensitive to the pose information. However, the image adversarial loss have much bigger effect on the unaligned dataset, which shows the adversarial loss can relief the strong constrain bring from image recon-

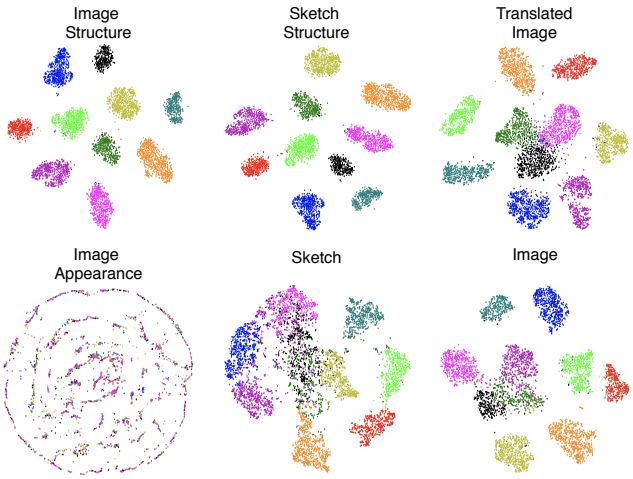

Figure 5: t-SNE results for different features on the testing set of Sketchy Ext. This figure is best viewed in color.

struction loss in unaligned data.

Alternation Analysis In the last two rows, we show two alternations, (1) move the orthogonality loss from $(f_{im}^{ap}, f_{im}^{st})$ to $(z_{im}^{ap}, f_{im}^{st})$ and (2) directly translate the image from the sketch structure feature without the image appearance feature z_{im}^{ap}. The performance drop after moving the orthogonality loss to z_{im}^{ap} and f_{im}^{st}, which indicate the effect of the disentangled representation. Besides, the performance also drop after we remove the variational estimator, which show the uncertainty compensation is crucial to train the image generator.

Retrieval Strategy Analysis In Figure 3, we plot the ω-P@200 curve. From this figure, we can find that the optimal ω is between 0.4 and 0.6. Besides, the peak of the curve usually close to the retrieval space who have higher performance.

4.4. Disentangle Analysis

To prove that our model actually disentangles the image features into structure space and appearance space, we first plot the varity of orthogonality loss and P@200 during the training and then we visualize the all the image and sketch featuresusing t-SNE [42].

As you can see in Figure 4, the orthogonality loss drops very slow at the first 6 training epoch. But during the 7-16 training epoch, the orthogonality loss drops very quickly. Compared with the variety of P@200, we find the similar tendency which indicates that the image disentanglement help our proposed model to generalize from seen to unseen categories in a great deal.

In Figure 5, we show the t-SNE results on image features, image appearance features, image structure features, sketch features, sketch structure features, sketch translated

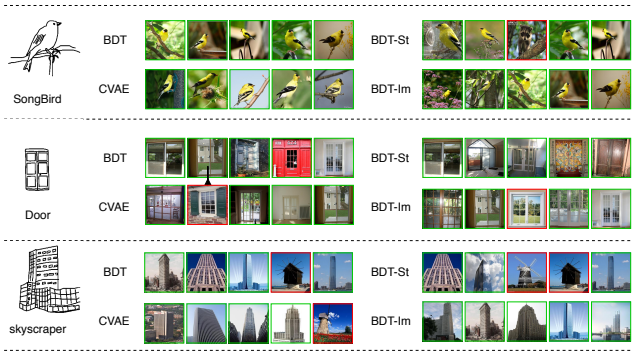

Figure 6: The top-5 images retrieved among BDT, BDT-St, BDT-Im, CVAE on Sketchy Ext. (Aligned) dataset. The green border indicates the right retrieval results and the red border indicates the opposite.

image features. From these figures we can find that the image feature is well disentangled into structure feature and appearance, in which the image appearance is cluster as different cycle in different appearance.Besides, the cluster results between image feature and translated image feature are similar and the cluster results between image structure feature and sketch structure feature are also similar, which also shows better alignment.

4.5. Case study

In Figure 6, we show the retrieval results of **BDT-St**, **BDT-Im**, **BDT** and CVAE [62]. Comparing them, we can find that the retrieval results in structure feature space concern more about the outlines and the retrieval results in image feature space care more about the pose and detailed features, which is similar to the results of CVAE, eg, for the sketch *door*, the retrieval results of both CVAE and BDT-Im have the *grid*, the results of BDT-St only contains the general outlines of *doorcase*.

5. Conclusion

This paper studies the problem of zero-shot sketch-based image retrieval from a new point of view, aka, using disentangled representation to facilitate the structure-based retrieval. To integrate the disentangled representation into a retrieval model, we proposed Bi-directional Domain Translation framework (BDT) in this paper and perform retrieval from two features space. Experiments on both Sketchy Ext. (aligned/unaligned) and TU-Berlin Ext. dataset demonstrate state-of-the-art performance. In further analysis and case study, we prove the model is capable of disentangling structure and appearance features for the image and the disentangled representation is indeed helpful for our model to generalize from seen categories to unseen categories.

References

[1] Xiaochun Cao, Hua Zhang, Si Liu, Xiaojie Guo, and Liang Lin. Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor. In *ICCV*, 2013. 1

[2] Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. Edgel index for large-scale sketch-based image search. In *CVPR*, 2011. 1

[3] Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *ACM MM*, 2010. 1

[4] Jiaxin Chen and Yi Fang. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In *ECCV*, 2018. 2

[5] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018. 2

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016. 2

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 3

[8] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 2

[9] Alberto Del Bimbo and Pietro Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):121–132, 1997. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 3, 5

[11] Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *NeurIPS*, 2018. 2, 6

[12] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 1, 2, 6, 7

[13] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? 2012. 5

[14] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 34(5):482–498, 2010. 1, 2

[15] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics*, 17(11):1624–1636, 2010. 1, 2

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 6, 7

[17] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *NeurIPS*, 2018. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 5

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 2

[20] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117(7):790–806, 2013. 1, 2

[21] Rui Hu, Tinghuai Wang, and John Collomosse. A bag-of-regions approach to sketch-based image retrieval. In *ICIP*, 2011. 1, 2

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 3

[24] Stuart James, Manuel J Fonseca, and John Collomosse. Reenact: Sketch based choreographic design from archival dance footage. In *ICMR*. 1

[25] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *ECCV*, 2018. 3

[26] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018. 2

[27] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 3

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014. 6

[29] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014. 2

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[31] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017. 6

[32] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 3

[33] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019. 3

[34] Ke Li, Kaiyue Pang, Yi-Zhe Song, Timothy Hospedales, Honggang Zhang, and Yichuan Hu. Fine-grained sketch-based image retrieval: The role of part-aware attributes. In *WACV*, 2016. 1

[35] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *NeurIPS*, 2018. 3

[36] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 2, 5, 6

[37] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, 2019. 2, 7

[38] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Multi-task adversarial network for disentangled feature learning. In *CVPR*, 2018. 2

[39] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *CVPR*, 2018. 3

[40] Peng Lu, Gao Huang, Yanwei Fu, Guodong Guo, and Hangyu Lin. Learning large euclidean margin for sketch-based image retrieval. *arXiv preprint arXiv:1812.04275*, 2018. 1, 7

[41] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 3

[42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9. 8

[43] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NeurIPS*, 2016. 3

[44] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 7

[45] George A Miller. *WordNet: An electronic lexical database*. 1998. 7

[46] Sarthak Parui and Anurag Mittal. Similarity-invariant sketch-based image retrieval in large databases. In *ECCV*, 2014. 1

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017. 6

[48] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, 2016. 1, 2

[49] Bernardino Romera and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 6

[50] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, 2015. 1, 2

[51] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016. 1, 2, 5

[52] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 1, 2, 5, 6

[53] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 3

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[55] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NeurIPS*, 2013. 6, 7

[56] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *CVPR*, 2017. 2

[57] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 2, 3

[58] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015. 1, 2, 6, 7

[59] Hao Wang, Cheng Deng, Xinxu Xu, Wei Liu, Xinbo Gao, and Dacheng Tao. Stacked semantic-guided network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:1904.01971*, 2019. 1, 2, 3

[60] Xinxun Xu, Hao Wang, Leida Li, and Cheng Deng. Semantic adversarial network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:1905.02327*, 2019. 1, 2, 6

[61] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2Image: Conditional image generation from visual attributes. In *ECCV*, 2016. 3

[62] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 1, 2, 3, 5, 6, 8

[63] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 2

[64] 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. *International journal of computer vision*, 122(3):411–425, 2017. 6

[65] Ruimao Zhang, Liang Lin, Rui Zhang, Wangmeng Zuo, and Lei Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Transactions on Image Processing*, 24(12):4766–4779, 2015. 6, 7

[66] Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017. 2

[67] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 2, 3

[68] Rong Zhou, Liuli Chen, and Liqing Zhang. Sketch-based image retrieval on a large scale database. In *ACM MM*, 2012. 1

[69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017. 3

[70] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 3