Style-Guided Zero-Shot Sketch-based Image Retrieval

Titir Dutta
titird@iisc.ac.in
Soma Biswas
somabiswas@iisc.ac.in

Department of Electrical Engineering
Indian Institute of Science
Bangalore, India

Abstract

Given a sketch query from a previously unseen category, the goal of zero-shot sketch-based image retrieval (ZS-SBIR) is to retrieve semantically meaningful images from a given database. The knowledge-gap between the seen and unseen categories along with sketch-image domain shift makes this an extremely challenging problem. In this work, we propose a novel framework which decomposes each image and sketch into its domain-independent content and a domain, as well as data-dependent variation/style component. Specifically, given a query sketch and a search set of images, we utilize the image specific styles to guide the generation of fake images using the query content to be used for retrieval. Extensive experiments on two large-scale sketch-image datasets, Sketchy extended and TU-Berlin show that the proposed approach performs better or comparable to the state-of-the-art in both ZS-SBIR and generalized ZS-SBIR protocols.

1 Introduction

Image retrieval is an important area of research in computer vision due to its potential applications in e-commerce, forensics etc. In sketch-based image retrieval (SBIR), hand-drawn sketches can be used as query to search in an image database, in case no natural image is available for a particular object. For example, if a user wishes to search for a specific dress in a fashion database, he/she can draw a free-hand sketch of that dress using a touch-screen device, since obtaining an image or writing a text-description of the same is difficult. The proliferation of touch-screen devices makes this problem very relevant in modern times.

Existing works on SBIR formulate the problem from two different perspectives. Fine-grained SBIR [32][18] refers to retrieval of the image instance which exactly matches the structure of the query sketch. Category based SBIR [19][36] deals in retrieving images of same category, but can be different in shape or pose from the sketch drawn by the user. However, for both the scenarios, during retrieval, the sketch query and the database images are assumed to belong to the categories used for training. For real-life applications, this assumption is quite restrictive, since objects of new categories are continuously being added to the database, which requires the model to be re-trained every time. Thus recently, zero-shot sketch-based image retrieval (ZS-SBIR) [29][38] is gaining increasing importance, where the query sketches and database images belong to categories, which are not seen during training. A further generalization over this protocol is proposed in [7], where a query from an unseen category is searched against a database of images from both seen and unseen categories. [38] demonstrates the degradation of retrieval performance of traditional SBIR methods for these protocols and hence justifies the need for further research in this direction. Few recent works have proposed fusion-based latent-space learning [29], generative model-based retrieval [38], semantic-aware cycle-consistent network [7] to address ZS-SBIR.

Inspired by the work in content-style disentanglement [3][12], to address the problem of ZS-SBIR, we propose to separate both the images and sketches into their domain-independent content and domain and data-specific variations/styles. The domain-independent content is a shared latent space for both domains [36][19], where the object representations follow a semantically meaningful order. During retrieval, this content-space is used to obtain the initial list of retrieved images. Motivated by the work in Zero-shot learning (ZSL) [21], which shows improved image classification performance in the image space as compared to the attribute space, we propose to use the image space for final retrieval. We further propose a cross-domain fusion of query content and search-set specific styles to obtain fake images, which are then used to get the final retrieved list. Specifically, given a query and its top-K retrieved images based on the content, we fuse the query content with the specific styles of each of these K-images to generate K-fake images. These fake images are finally compared with the corresponding search set images to get the final retrieved images. The contribution of this work is summarized here

- We propose a style-guided image generation during retrieval to eliminate the effect of domain difference and intra-class variations for improved performance.
- We effectively utilize the concepts of content-style separation for the task of ZS-SBIR, by separating the original data representations into a semantic-aware domain-invariant content and domain and data specific variations/style.
- Experiments on two large-scale sketch-image datasets are conducted and comparisons with the state-of-the-art is reported.

The rest of the paper is organized as follows. Section 2 gives a brief description of the related work. Our proposed approach is discussed in Section 3, followed by extensive experiments and analysis in Section 4. Finally, we conclude in Section 5.

2 Related work

In this section, we briefly review the current literature in the field of SBIR, ZSL, ZS-SBIR as well as the content-style disentaglement methods.

Sketch-based Image retrieval (SBIR): The main challenge in SBIR is the domain-gap between the sketch and image representations. Early methods using hand-crafted features attempt to bridge the domain gap by using edge-maps extracted from images. Then, specially designed features, HOG [14], LKS [27], etc. are used to represent both the edge-maps and sketches and retrieval is performed by comparing similarity between the features. In contrast, the deep-learning based methods learn some variants of end-to-end cross-modal retrieval models to address this sketch-to-image domain shift. Some of the recent methods are siamese networks [24], triplet-loss [28] or contrastive-loss [4] based models. [2] describes a hybrid multi-stage deep network, which combines both contrastive and triplet networks. Recently, Deep Sketch Hashing (DSH) [19] proposed a hashing-based heterogeneous network, which in addition to deep-features of sketch and images, uses the edge-maps extracted from images as sketch-tokens to gain state-of-the-art retrieval accuracy. A non-deep shared-space learning method [36] expoliting curriculum learning has also been proposed. [5] adddresses a variant of SBIR, where the preferred aesthetic style of the retrieved images is specified as an additional constraint with the sketch query.

Zero-shot Sketch-based Image Retrieval (ZS-SBIR): To evaluate the generalizability to novel class data, in ZS-SBIR [29][38][7], the sketch query and the image database contain samples from categories which are different from the training categories. In [29], a sketch-image feature fusion-based end-to-end model has been trained to perform retrieval. While the performance is reasonable, still the fusion architecture has a high memory-requirement, as identified in [7]. A sketch-to-image feature generation and voting-based retrieval method in the image-space has been proposed in [38]. However, this approach requires paired sketch-image training data, which may be difficult to obtain in practice. A cycle-consistent loss-based semantically-aligned latent-space has been proposed in [7]. This model involves learning two parallel GAN-models for sketches and images respectively. [6] proposes a new large-scale sketch-image dataset for SBIR and proposed a triplet loss-based network.

Zero-shot learning (ZSL): A related research area, ZSL addresses the problem of classifying images which belong to classes unknown to a trained classifier. Initial ZSL methods propose to learn a latent embedding space [16][25], which contains the semantic information (attribute) of training data and while testing, the representation of images in this latent space predicts the target category. An alternate approach [21][22][34] of synthesizing image features from such attributes have recently achieved high performance.

Content-Style disentanglement: Our work is motivated by the success of the disentanglement methods applied to several computer vision applications, such as, style transfer [11], image-to-image translation [15][20]. Recently, a number of disentanglement algorithms [3], [12] have been proposed, where the objective is to separate the identity, which is important from the classification perspective and the style, which holds additional information about the image, but not important for classification. But there are significant differences between our content-style decomposition and the existing literature, as will be discussed later and to the best of our knowledge, this is the first work, which uses this concept for ZS-SBIR.

3 Approach

In this section, we explain the proposed method in details. Given the query sketch and a database image, during retrieval, we propose to generate a fake image with the content same as query and style of the database image, which will finally be compared with the database image. This reduces the domain difference and the image-specific variations, which results in improved retrieval performance. To achieve this goal, we need two modules, namely the (1) content-style decomposition module and (2) content-style fusion module, which we will describe in details in the following sub-sections. The training is performed in two phases. The first module learns to decompose the sketch/image to obtain the domain-invariant shared space *content*, and domain and data-dependent *style* representations. The second module learns to fuse the query content and image-specific styles to generate fake image features to obtain the final retrieved images. Figure 1 illustrates the proposed approach.

Notations: Let us denote the available sketch and image data for training as $\mathcal{S}_{train} = \{\mathbf{S}_i, l_i^{(S)}\}_{i=1}^{N_S}$ and $\mathcal{I}_{train} = \{\mathbf{I}_i, l_i^{(I)}\}_{i=1}^{N_I}$ respectively. $l_i^{(S)}$ (or $l_i^{(I)}$) represents the label of i^{th} sketch $\mathbf{S}_i$ (or image $\mathbf{I}_i$) and $l_i^{(S)}$, $l_i^{(I)} \in \mathcal{Y}_{seen}$, for all i, where $\mathcal{Y}_{seen}$ represents the set of all

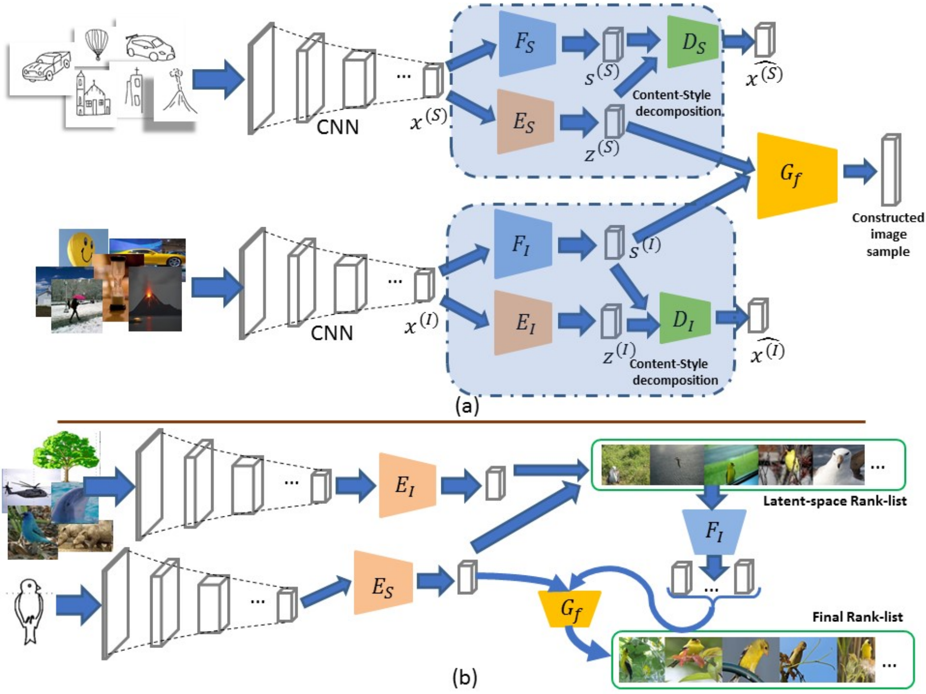

Figure 1: (a) Illustration of the proposed framework for (a) Training and (b) Retrieval.

classes available for training (seen classes). In contrast to the existing literature, no pairing [38] or same index [7] assumptions of sketch and image data are considered for training in our work. The testing sketch and image data are denoted as, $\mathcal{D}_{sketch} = \{\mathbf{S}_j, l_j^{(S)}\}_{j=1}^{M_S}$ and $\mathcal{D}_{image} = \{\mathbf{I}_j, l_j^{(I)}\}_{j=1}^{M_I}$ respectively. In this case, $l_j^{(S)}$, $l_j^{(I)} \in \mathcal{Y}_{unseen}$ and $\mathcal{Y}_{unseen} \cap \mathcal{Y}_{seen} = \phi$.

Feature representation: First, the features of the images and sketches are obtained by fine-tuning a pre-trained convolutional network (CNN) using $\mathcal{S}_{train}$ and $\mathcal{I}_{train}$ separately. The output of the penultimate layer of each fine-tuned network, $\mathbf{x}_i^{(S)}$ and $\mathbf{x}_i^{(I)}$ are considered as initial representations of the sketch and image respectively. Since, the CNN is trained on large-scale image dataset with label-based cross-entropy loss for classification [30], the extracted features ideally should contain only the category information of the image. But since they are trained on images and fine-tuned on images and sketches separately, they still contain the domain specific information. We use the content-style decomposition module to decompose these features into the domain-independent *content* representation and the *style*, which contains the domain-dependent part and the residual data-specific information. Once fine-tuned, these network weights are kept fixed for the rest of the training.

3.1 Content-Style decomposition module

To generate style-guided image features for retrieval, first we learn to separate the content-style information of sketch/ image features in a two stage adversarial fashion. Though this decomposition is inspired by [12], there are some important differences which are explained later. Specifically, we learn two content encoders E_S and E_I for sketches and images respectively, to encode the *content* information as $\mathbf{z}_i^{(m)} = E_m(\mathbf{x}_i^{(m)})$, where $m \in \{S, I\}$. In addition, the content extracted from sketches and images of the same categories should be the same. As in ZSL [35], to ensure better generalization of the model to unseen categories, we enforce these shared latent representations to be similar to their respective category-name embeddings, $h(l_i^{(m)}), m \in \{S, I\}$. Thus, given a sketch $\mathbf{S}$ (or image $\mathbf{I}$), the probability that the sample belongs to the category $l^{(S)}$ (or $l^{(I)}$) in the shared space is measured as

$$p(l^{(m)}|\mathbf{x}^{(m)}) = \frac{e^{-\alpha d(\mathbf{z}^{(m)}, h(l^{(m)}))}}{\sum\limits_{l^{(m)} \in \mathcal{Y}_{seen}} e^{-\alpha d(\mathbf{z}^{(m)}, h(l^{(m)}))}}, \qquad m \in \{S, I\} \tag{1}$$

where, $d(\mathbf{z}^{(m)}, h(l^{(m)})) = ||\mathbf{z}^{(m)} - h(l^{(m)})||^2$ represents the distance between the extracted content and the ground-truth category-name embedding. Thus, the latent-space representations are learned with the distance-based cross-entropy loss on E_S and E_I as,

$$\mathcal{L}_m = -logp(l^{(m)}|\mathbf{x}^{(m)}), \qquad m \in \{S, I\} \tag{2}$$

Such category-name embedding guided learning of latent-space representations follow a semantically meaningful structure, which forces similar class samples from both domains to be closer to each other. This helps to project unseen class sketch/image samples close to semantically similar seen classes and thus, the distance between unseen class sketch/image representations of the same category is minimized.

In the second stage of learning the decomposition, we capture the styles present in $\mathbf{x}^{(m)}, m \in \{S, I\}$. Following [12], we learn two style encoders F_S and F_I for sketches and images, such that the learned styles $\mathbf{s}^{(m)}, m \in \{S, I\}$ contain all domain and data dependent information, which are not useful from recognition perspective. Additionally, we learn decoder networks $D_m(\mathbf{z}^{(m)}, \mathbf{s}^{(m)}), m \in \{S, I\}$, such that given the content $\mathbf{z}^{(m)}$ and style $\mathbf{s}^{(m)}$, it can reconstruct back the corresponding sketch or image features, $\mathbf{x}^{(m)}$. Hence, in this stage, we minimize the following joint loss function

$$\mathcal{L}^{style} = \mathcal{L}_m^{rec} - \gamma \mathcal{L}_m^{adv} \tag{3}$$

where, the reconstruction loss component is defined as, $\mathcal{L}_m^{rec} = ||\mathbf{x}^{(m)} - D_m(\mathbf{z}^{(m)}, \mathbf{s}^{(m)})||^2$. The adversarial loss $\mathcal{L}_m^{adv} = -logp_{adv}(l^{(m)}|\mathbf{x}^{(m)})$ ensures that the style features $\mathbf{s}^{(m)}$ does not contain any useful information. In this case, $p_{adv}(l^{(m)}|\mathbf{x}^{(m)})$ is measured as *softmax(*$\mathbf{s}^{(m)}$*)*. γ is a hyper-parameter which is set experimentally. In our implementation, the encoders consist of two fully-connected (fc) layers with ReLU activations. The decoder networks consist of a concatenation layer followed by two fc layers with ReLU-activations.

3.2 Content-Style fusion module

Our final goal is to generate style-guided image using the query content and search-set image styles, which requires a content-style fusion module. Here, we learn a concatenation-based fusion network G_f, which can combine cross-domain content-style features to construct meaningful image features. For training this module, we fix the first part of the network, i.e., the encoders, adversarial classifiers and the decoders. To train G_f, we construct a triplet set $\mathcal{T} = \{\mathbf{S}_i, \mathbf{I}_i^+, \mathbf{I}_i^-\}_{i=1}^N$ from $\mathcal{S}_{train}$ and $\mathcal{I}_{train}$, such that, $l_i^{(S)} = l_i^{(I^+)}$ and $l_i^{(S)} \neq l_i^{(I^-)}$. The constructed image features $\hat{\mathbf{x}}^{(I)} = G_f(\mathbf{z}^{(S)}, \mathbf{s}^{(I)})$ are restricted to follow a margin-based categorical similarity by the following triplet-ranking loss function as

$$\mathcal{L}_I^{triplet} = \sum_{i=1}^{N} \max\ \{0, [d(\hat{\mathbf{x}}_i^{(I)}, \mathbf{x}_i^{(I^+)}) - d(\hat{\mathbf{x}}_i^{(I)}, \mathbf{x}_i^{(I^-)}) + M]\} \tag{4}$$

where M is the margin, set experimentally. Here, G_f learns to construct image features from a sketch content by fusing it with image-specific style information, such that the generated fake image feature comes closer to other image features from the same class (image matching problem). In our design, the fusion module contains a concatenation layer followed by two fc layers with ReLU activation.

3.3 Retrieval methodology

Category-based SBIR approaches are evaluated based on whether the category of the input query sketch matches with that of the retrieved images [29][38]. During retrieval, given a sketch query and a set of database images, their feature representations are passed through the content-style decomposition module to obtain their content and style representations. First, we perform the retrieval solely based on their content in the shared latent space and create a complete ranked list using ascending Euclidean distance between the query sketch content and search set image contents. Furthermore, we select top-K retrieved images against a sketch query in the latent-space as our initial pruned rank-list $\mathcal{R}_{latent}$. Fusing the styles of these top-K images with the query sketch content, we generate K-fake image features. The intuition is that since this step eliminates the variations due to domain and specific data samples, if the content of the query matches with that of a search set image, the distance between that search set image and its corresponding fake image will be very small. Though, it may seem that this is the same as content matching, it has been shown extensively in ZSL literature that matching in the image domain has significant advantages over matching in the attribute space [21]. Even in our experiments, we obtain significant performance difference between the two approaches as will be shown later. In this image space, we compute the Euclidean distance between the i^{th} generated fake image $\hat{\mathbf{x}}_i^{(I)}$ and the original image $\mathbf{x}_i^{(I)}$ in $\mathcal{R}_{latent}$ and consider that to be the final distance computed between the query sketch and the i^{th} image. On the basis of this newly calculated distances, we obtain the final rank-list as $\mathcal{R}_{image}$, which is in effect a re-ranked version of top-K images in the latent-space.

3.4 Difference with Existing Work

Difference with [38]: CVAE and CAAE are two generative model based ZS-SBIR methods, which propose to generate image features from a given sketch for retrieval. Although conceptually similar, our method differs from CVAE in a number of ways; 1) CVAE requires paired image-sketch data to learn the generation, while the proposed method has no such restriction; 2) Our method fuses image-specific styles with sketch-content and uses the similarity with these fake images for retrieval. In contrast, CVAE generates samples from Gaussian noise and employs a voting methodology for retrieval.
Difference with [12]: In [12], a general content-style disentanglement technique is proposed, whereas we specifically employ the decomposition with the final goal of retrieval. For the same, we use a modified distance-based cross-entropy loss in contrast to the softmax probability based classification loss, as in [12]. Additionally, the *style*-definition in [12]

is significantly different from ours. We encode the domain-specific knowledge, as well as intra-class variations in the style-vectors, whereas [12] encodes the class-independent information as style. In our work, we deal with features extracted from fine-tuned networks and not the original image data as in [12].

4 Experiments

Here, we present the results of the experiments performed to evaluate the effectiveness of the proposed approach. First, we provide a brief description of the dataset used in this work.

The Sketchy Dataset [28] is a collection of 75,471 sketches and 12,500 images from 125 classes. For our experiments, we used the extended dataset [19], which contains additional 60,502 images. For ZS-SBIR, two different data-splits have been proposed in literature. In the first split (**Split 1**), randomly chosen 25 classes are considered as unseen and rest 100 classes are used for training. This split is followed by [29][7] and they utilize pre-trained AlexNet [17] and VGG-16 [30] features for image/sketch representations. Since some of the classes in this dataset are part of ImageNet [26], hence it is possible that pre-trained AlexNet or VGG-16 has already been trained with few of these unseen classes. However, the training method in [29][7] ensures that the mapping between unseen-class sketch-image are not known by the model. In contrast, [38] propose another carefully designed data-split (**Split 2**), where 21 categories, which are not part of the ImageNet [26] are selected as unseen and the rest are used for training. This split ensures that the unseen classes are truly unknown to the retrieval system.

TU-Berlin dataset [8] contains 20,000-sketches from 250 categories and is extended by [19] with 2,04,489 natural images. A random split of 220 training classes and 30 testing classes with at least 400 images per category are used in literature [29][7] for ZS-SBIR.

Implementation Details: The proposed model is implemented using TensorFlow [1]. All hyper-parameters are tuned based on the accuracy on validation set, constructed as 10% of the training set. We fine-tuned pre-trained VGG-16 separately for images and sketches using our training data and the fc7-features are considered as $\mathbf{x}^{(m)}, m \in \{S, I\}$. The category-name embeddings $h(.)$ have been computed as 200-d vectors using a pre-trained GloVe [23] model. The dimensions of $\mathbf{z}^{(m)}$ and $\mathbf{s}^{(m)}, m \in \{S, I\}$ are restricted to 200-d and 100-d, respectively. Adam optimizer has been used for optimization with $\beta_1 = 0.5, \beta_2 = 0.999$ and a learning rate of 10^{-3} for the first phase of content-style decomposition and 10^{-4} for the second phase of fusion with a batch-size of 64 and 32 for Sketchy and TU-Berlin, respectively.

Performance comparison: We compare our method with several state-of-the-art SBIR methods, as well as ZS-SBIR algorithms. To compare directly with results reported in the literature, we perform experiments on both Split 1 and Split 2 for Sketchy and the standard split for TU-Berlin using the same evaluation metric as in the respective papers.

Table 1 reports the results on Sketchy (Split 1) and TU-Berlin (standard ZS-SBIR split), used in [29][7] in terms of MAP@all and Precision@100. All the results for the other approaches are directly taken from [7]. We observe that proposed method significantly outperforms all state-of-the-arts using both the metrics on Sketchy. For TU-Berlin, we achieve second best performance, which is only less than [7] and better than all the others. Follow-

Type	Methods	Sketchy (Split 1)		TU-Berlin	
		Precision@100	MAP@all	Precision@100	MAP@all
SBIR	Softmax Baseline	0.172	0.114	0.143	0.089
	Siamese CNN [24]	0.175	0.132	0.141	0.109
	SaN [39]	0.125	0.115	0.108	0.089
	GN Triplet [28]	0.296	0.204	0.253	0.175
	3D shape [33]	0.078	0.067	0.067	0.054
	DSH [19]	0.231	0.171	0.189	0.129
	GDH [40]	0.259	0.187	0.212	0.135
ZSL	CMT [31]	0.102	0.087	0.078	0.062
	DeViSE [10]	0.077	0.067	0.071	0.059
	SSE [41]	0.161	0.116	0.121	0.089
	JLSE [42]	0.185	0.131	0.155	0.109
	SAE [16]	0.293	0.216	0.221	0.167
	FRWGAN [9]	0.169	0.127	0.157	0.110
	ZSH [37]	0.214	0.159	0.177	0.141
ZS-SBIR	ZSIH [29]	0.342	0.258	0.294	0.223
	ZS-SBIR [38]	0.284	0.196	0.001	0.005
	SEM-PCYC [7]	0.463	0.349	0.426	0.297
	Proposed	**0.4842**	**0.3756**	**0.3551**	**0.2543**
Generalized ZS-SBIR	ZSIH [29]	0.296	0.219	0.218	0.142
	SEM-PCYC [7]	0.364	0.307	0.298	0.192
	Proposed	**0.3811**	**0.3307**	**0.2264**	**0.1488**

Table 1: Comparison (Precision@100 and MAP@all) of the proposed method with state-of-the-art ZS-SBIR methods on Sketchy (Split 1) and TU-Berlin datasets.

ing [7], we also report the performance for the generalized ZS-SBIR protocol, where both seen and unseen class images are present in the search set. For this protocol also, we achieve the best and second best performance on the two datasets, compared to the state-of-the-art.

We perform additional experiments on Sketchy Split 2. Table 2 reports MAP@200 and Precision@200 for the same. As before, the results of the other approaches are directly taken from [38]. Our search-set specific style-guided retrieval clearly outperforms all approaches, including both noise-based generation methods CVAE and CAAE, by significant margins.

The top-5 retrieved images for few randomly sampled sketches from the test set in Sketchy Split 2 are shown in Figure 2. In Figure 2(a), we show examples for which all 5 retrieved images are correct, while in Figure 2(b), some of the retrieved images are incorrect. Here we make an interesting observation; the incorrect retrievals are not from completely random categories, instead, they share a considerable semantic similarity with the query sketches. For example, *cow*-images are retrieved incorrectly against a query sketch of *rhinoceros* (both are animals), *sword* is retrieved incorrectly for *scissors* (both are tools).

Analysis of Proposed Framework: Here, we analyze our framework by developing different baselines by modifying individual modules or the loss-terms. The results obtained are reported in Table 3 on Sketchy Split 2 [38]. In B1, the images are retrieved using only the content of sketches $(\mathbf{z}^{(S)})$ and images $(\mathbf{z}^{(I)})$ in $\mathcal{R}_{latent}$. B2 reports the results when the content-encoder is trained using standard softmax cross-entropy loss using class labels instead of category-word vectors. Since such loss function does not contain any semantic information about the categories, the unseen class retrieval accuracy is lower. In B3, we perform a fusion of the content and image space similarity scores to get the final retrieval results. However, to our surprise, this did not give any improvement over the image-space retrieval. This may be because the fake images are constructed by adding style information

Type	Evaluation methods	Precision@200	MAP@200
SBIR methods	Baseline	0.106	0.054
	Siamese-1 [13]	0.243	0.134
	Siamese-2 [4]	0.251	0.149
	Coarse-grained Triplet [28]	0.169	0.083
	Fine-grained Triplet	0.155	0.081
	DSH [19]	0.153	0.059
ZSL methods	Direct Regression	0.066	0.022
	ESZSL [25]	0.187	0.117
	SAE [16]	0.238	0.136
ZS-SBIR	CAAE [38]	0.260	0.156
	CVAE [38]	0.333	0.225
	Proposed	**0.4001**	**0.3581**

Table 2: Comparison (Precision@200 and MAP@200) of the proposed method with existing ZS-SBIR methods on Split 2 of Sketchy data.

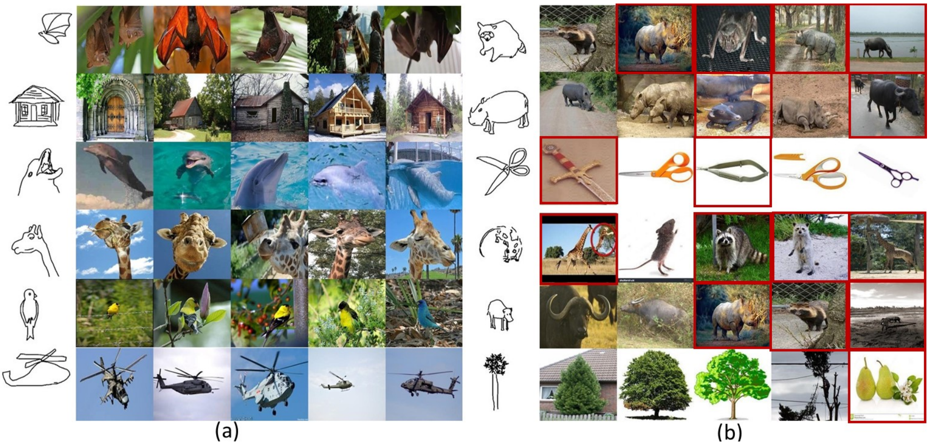

Figure 2: Top-5 retrieved images for few sketch queries from Sketchy data - Split 2. (a) All retrieved images are correct; (b) Few retrieved images (in *red*-frame) are incorrect.

to the sketch content, and hence it may not contain any complimentary information to boost the performance. Finally, in B4, we generated a single fake image from the query by fusing it with a randomly selected style from the database. We perform retrieval on the basis of similarity of this single generated image with all the images in search set. With such an approach, we observe a considerable drop in performance, which justifies the effectiveness of appropriate modelling of styles and image-specific style fusion used in our framework. The full proposed model, with style-based final ranking, produces the best result.

5 Conclusion

In this work, we proposed a content-style decomposition-based ZS-SBIR model, which either outperforms or yields comparable results with the state-of-the-art algorithms for both ZS-SBIR and generalized ZS-SBIR protocols. We observe an improved generalization ability of the proposed retrieval system by exploiting the domain-specific as well as image-specific style information of the database images. Extensive experiments have been per-

Description	MAP@200
B1: Content-based retrieval in the shared space	0.3280
B2: Label-based CE-loss for content-style decomposition	0.3228
B3: Score-fusion	0.3213
B4: Single fake-image (with random style) based retrieval	0.2854
Proposed model	**0.3581**

Table 3: Comparison of the proposed framework with different baselines.

formed on two recent large-scale sketch-image datasets. The improved performance using our framework on multiple data splits shows the effectiveness of the content-style separation in the context of cross-domain data retrieval.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*, 2016.

[2] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse. Sketching out the details: sketch-based image retrieval using convolutional neural networks with multi-stage regression. *C & G*, 71:77–87, 2018.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: interpretable representation learning by information maximizing generative adversarial nets. *arXiv:1606.03657v1*, 2016.

[4] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[5] J. Collomosse, T. Bui, M. Wilber, C. Fang, and H. Jin. Sketching with style: visual search with sketches and aesthetic context. In *ICCV*, 2017.

[6] S. Dey, P. Riba, A. Dutta, J. Llados, and Y. Z. Song. Doodle to search: practical zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[7] A. Dutta and Z. Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[8] E. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM TOG*, 31(4): 1–10, 2012.

[9] R. Felix, B. G. Vijay Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018.

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: a deep visual-semantic embedding model. In *NIPS*, 2013.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[12] N. Hadad, L. Wolf, and M. Shahar. A two-step disentanglement method. In *CVPR*, 2018.

[13] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

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

[15] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[16] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[18] Y. Li, T. M. Hospedales, Y. Z. Song, and S. Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014.

[19] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: fast free-hand sketch-based image retrieval. In *CVPR*, 2017.

[20] M. Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Neur-IPS*, 2017.

[21] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han. From zero-shot learning to conventional supervised classification: unseen visual data synthesis. In *CVPR*, 2017.

[22] A. Mishra, M. S. K. Reddy, A. Mittal, and H. A. Murthy. A generative model for zero-shot learning using conditional variational autoencoders. In *CVPRW*, 2018.

[23] J. Pennington, R. Socher, and C. D. Manning. Glove: global vectors for word representation. In *EMNLP*, 2014.

[24] Y. Qi, Y. Z. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, 2016.

[25] B. Romera-Paredes and P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. F. Li. Imagenet: large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[27] J. M. Saavedra and J. M. Barrios. Sketch-based image retrieval using learned keyshapes (lks). In *BMVC*, 2015.

[28] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 35(4):1–12, 2016.

[29] Y. Shen, L. Liu, F. Shen, and L. Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[31] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[32] 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 *ICCV*, 2017.

[33] M. Wang, C. Wang, J. X. Yu, and J. Zhang. Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. *VLDB*, 8(10): 998–1009, 2015.

[34] 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. In *AAAI*, 2018.

[35] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018.

[36] D. Xu, X. Alameda-Pineda, J. Song, E. Ricci, and N. Sebe. Cross-paced representation learning with partial curricula for sketch-based image retrieval. *IEEE T-IP*, 27(9):4410–4421, 2018.

[37] Z. Yang, W. W. Cohen, and R. Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, 2016.

[38] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal. A zero-shot framework for sketch-based image retrieval. In *ECCV*, 2018.

[39] 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. *IJCV*, 122(3):411–425, 2017.

[40] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. Tao Shen, and L. Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, 2018.

[41] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE T-IP*, 24(12):4766–4779, 2015.

[42] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.