

s-SBIR: Style Augmented Sketch based Image Retrieval

Titir Dutta

Indian Institute of Science, Bangalore

titird@iisc.ac.in

Soma Biswas

Indian Institute of Science, Bangalore

somabiswas@iisc.ac.in

Abstract

Sketch-based image retrieval (SBIR) is gaining increasing popularity because of its flexibility to search natural images using unrestricted hand-drawn sketch query. Here, we address a related, but relatively unexplored problem, where the users can also specify their preferred styles of the images they want to retrieve, e.g., color, shape, etc., as keywords, whose information is not present in the sketch. The contribution of this work is three-fold. First, we propose a deep network for the problem of style-augmented SBIR (or s-SBIR) having three main components - category module, style module and mixer module, which are trained in an end-to-end manner. Second, we propose a quintuplet loss, which takes into consideration both the category and style, while giving appropriate importance to the two components. Third, we propose a normalized composite evaluation metric or ncMAP which can quantitatively evaluate s-SBIR approaches. Extensive experiments on subsets of two benchmark image-sketch datasets, Sketchy and TU-Berlin show the effectiveness of the proposed approach.

1. Introduction

Sketch-based image retrieval (SBIR) addresses the problem of retrieving semantically relevant natural images from a search set, given a roughly-drawn sketch as query. This research direction is gaining increasing attention as a sketch has the potential to reflect the search-requirement better compared to describing it in texts [50]. But, free-hand sketches can be quite different compared to natural images, since humans tend to focus mainly on the object structure and not on the finer details. Thus, the main challenge for the retrieval algorithms is to bridge the significant domain gap between the sketch query and the database images. Many recent approaches derive a shared feature-space to address SBIR [29][44][25].

Here, we address a related, but relatively unexplored problem, where, while searching for a particular object category, e.g. *car*-image using a sketch, the user also has the flexibility to specify their preferred styles/attributes in the

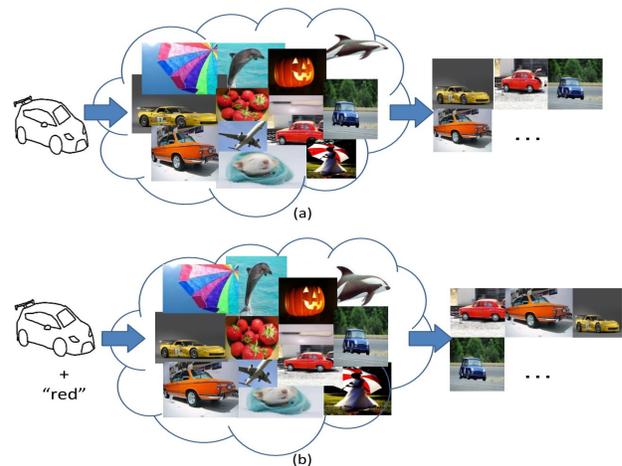


Figure 1. (a) In SBIR, given a sketch query, the goal is to retrieve images belonging to the same category. (b) In s-SBIR, the user can also give one/multiple desired styles, and the goal is to retrieve relevant images of the correct category having the desired styles.

form of ‘keywords’ (e.g., *red* in color) (Figure 1). Recently, researchers have started to explore this problem in the single-domain scenario (image-to-image retrieval) [24][49]. A variation of the s-SBIR problem is addressed in [6], but there are significant differences, which we will discuss later.

To address style-augmented SBIR (s-SBIR), we propose a novel deep-framework which analyzes the category and style information of the images appropriately to retrieve relevant images based on the hand-drawn sketch category and the user-specified styles. The framework consists of (1) *content module* to extract the content of the images and sketches; (2) *style module* to extract the style of the images such that they match with the desired user-specified styles encoded in the keywords and (3) *mixer module* which combines the sketch content and styles from the keywords such that the relevant images can be directly retrieved. The contributions of this work can be summarized as follows

- We propose a variant of the standard SBIR problem, where the user-required styles can be specified in terms of ‘keywords’.

- The unequal importance of category and style is incorporated in a quintuplet loss function conditioned on both the factors for the final retrieval.
- For s-SBIR evaluation, we propose normalized composite Mean Average Precision (ncMAP), a suitable modification of the standard evaluation metric MAP.
- Extensive experiments on two SBIR data-subsets, namely m-Sketchy [34] and m-TU-Berlin [8] demonstrate the effectiveness of proposed framework.

The rest of the paper is organized as follows. The related work is discussed in Section 2. The proposed method is described in Section 3 followed by the experiments and discussions in Section 4. The paper ends with a conclusion.

2. Related Work

In this section, we provide pointers to some of the recent work in literature and how our work differs from them.

Sketch-based Image Retrieval: Previously, most of the methods for SBIR used low-level features, such as color histogram [18], elastic contours [2], as well as local features like SIFT [26], etc. for both sketches and images and the retrieval was performed based on feature-matching. A mid-level or common-domain feature learning algorithm is introduced in [30][33]. [9] uses the concatenated Bag-of-Words (BoW) features to be the common-domain representation. Using the edge information of images, a perceptual grouping framework has been introduced in [30].

Recently, hashing-based cross-modal retrieval has become very popular and has been applied to SBIR as well. However, in hashing-based image-retrieval, the image search set remains exactly the same for both training and testing. [3][10] are some of the recent literature to report efficient hashing methods for SBIR. A heterogeneous neural network based hashing method is proposed in [25], which achieves the current state-of-the-art for SBIR on two large-scale sketch-image datasets. This work integrates a deep-network based image/sketch representation learning with a non-deep framework for measuring their similarity. A cross-paced curriculum based dictionary learning technique has been introduced in [43]. A clustering-based re-ranking approach is proposed in [15]. Note that standard SBIR deals in category-based retrieval and does not have any provision to incorporate the styles specified by the user.

Attribute-based Image-to-Image Retrieval: Image-to-image retrieval using one or more styles/attributes as additional input has gained popularity in recent literature. [24] proposes a hashing based method where the binary attribute vector can be used as query to retrieve required images. The fashion-image search framework AMNet [49] retrieves images from fashion items gallery, based on a query image (a dress) and an user-defined style (*pink* in color). Re-

cently, [39] studies the image-retrieval problem, where in addition to the search image, a text query describing the desired modification in the retrieved image is used. All these approaches address the attribute-based retrieval problem in single-domain and the results are qualitative.

Fine-Grained Sketch-based Image Retrieval: Here the goal is to retrieve images from the search set based on the object category as well as subtle appearance details drawn in the sketch [21][44][37]. The high-level objective of fine-grained SBIR (FG-SBIR) is somewhat similar to the proposed s-SBIR in the sense that both specify additional criteria on the retrieval algorithm apart from just the category match. Note that the proposed framework can also be integrated with FG-SBIR to specify styles like color, texture or material, which can not be done using the sketch, though in this work, we focus on generic SBIR.

Content-Style Disentangled Representation Learning: We take motivation from the literature in learning disentangled representations [5][12] to separate the content and style information of the input data. Although conceptually similar approaches have been used for image-to-image translation [16], style transfer [11], this is the first time such an approach is effectively used for SBIR.

Sketch-based Image retrieval with style: Recently, [6] proposed a variant of SBIR, where in addition to the sketch, another image is used as query. This additional image holds the aesthetic / style requirement of the user. Even though similar, s-SBIR problem is significantly different from the one addressed in this paper.

3. Proposed Approach

In this section, we formalize the problem statement and discuss in details the proposed approach. Let the image data be denoted as $\mathcal{D}_{\mathcal{I}} = \{\mathbf{I}_i, \mathbf{y}_i^I, \mathbf{a}_i^I\}_{i=1}^{n_I}$, where \mathbf{I}_i is the i^{th} image and \mathbf{y}_i^I is the corresponding label vector. $\mathbf{a}_i^I \in \mathbb{Z}^{d_a}$ represents the ground-truth style information, example, color, shape, texture, etc. Assuming there are d_a number of annotated styles, an entry of 1 in \mathbf{a}_i^I indicates the presence of that style, otherwise it is 0. n_I is the total number of images available for training. The sketch dataset consisting of n_S samples is denoted as $\mathcal{D}_S = \{\mathbf{S}_i, \mathbf{y}_i^S\}_{i=1}^{n_S}$.

The proposed framework (Figure 2) has three main modules, (1) category module; (2) style module and (3) mixer module. Here, we describe all the modules and their functionalities. The network details are provided in Section 4.

3.1. Category Module

In this module, given an image and sketch, the goal is to learn a shared representation based solely on their category. This consists of a feature generation part, followed by an encoder-decoder structure for obtaining the same.

Feature Generation: For extracting features which can capture the input content, we utilize a pre-trained classi-

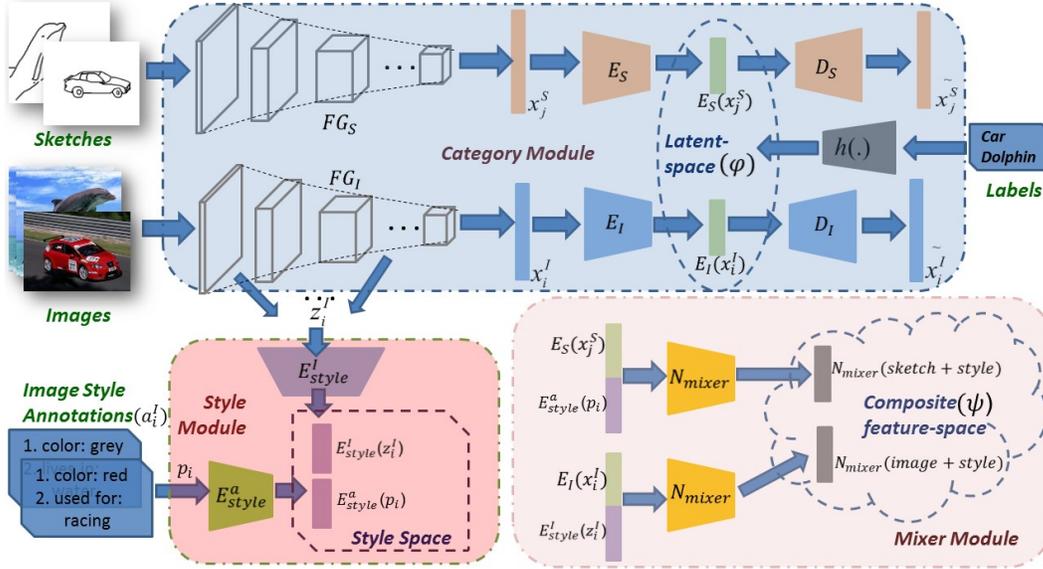


Figure 2. An illustration of the proposed end-to-end framework for s-SBIR. This figure is best viewed in color.

fication network, since the features extracted from such a model is trained to capture only the category information of the input. In this work, we fine-tune the pre-trained VGG-19 [36] (originally trained to perform object classification [36] on ImageNet [31]) on images and sketches separately for our datasets. Such fine-tuned networks are denoted as FG_I and FG_S , respectively. The output of the last fully-connected layer ($fc7$) from both networks, $\mathbf{x}_i^I \in \mathbb{R}^{d_I}$ and $\mathbf{x}_j^S \in \mathbb{R}^{d_S}$ are considered as the feature representations of \mathbf{I}_i and \mathbf{S}_j respectively in their own domain. After fine-tuning, the weights of FG_I and FG_S are frozen for the remaining part of network training.

Latent-space representation: The generated features \mathbf{x}_i^I and \mathbf{x}_j^S are domain specific and thus cannot be compared directly. Instead, we learn a shared latent Φ -space for direct comparison. Here, we learn a multi-layer perceptron (MLP)-based encoder-decoder network to obtain the Φ -space representations. We want to construct this space such that (1) the samples from same categories of both domains are close and (2) the samples from different categories are far apart while maintaining their semantic relation (semantically similar classes, e.g., *car* and *bus* should be closer compared to semantically dissimilar classes, e.g. *car* and *umbrella*). Such a Φ -space structure is achieved using a pre-trained word-embedding model GloVe [28].

Latent-space loss: To transform the features \mathbf{x}_i^I and \mathbf{x}_j^S to the Φ -space, two MLP-encoders E_I (for image) and E_S (for sketch) are designed and the following loss is minimized,

$$\mathbb{L}_{latent} = \sum_{m \in \{I, S\}} \sum_{i=1}^{n_m} \|E_m(\mathbf{x}_i^m) - h(\mathbf{y}_i^m)\|^2 \quad (1)$$

where $h(\mathbf{y}_i^m) \in \mathbb{R}^d$ represents the GloVe-embedding vector representation of the category-name of the i^{th} sample.

Reconstruction loss: We design two MLP-decoders D_I and D_S for images and sketches respectively to enforce their shared representations to retain the necessary details for reconstruction [19]. The loss used is given as

$$\mathbb{L}_{recons} = \sum_{m \in \{I, S\}} \sum_{i=1}^{n_m} \|D_m(E_m(\mathbf{x}_i^m)) - \mathbf{x}_i^m\|^2 \quad (2)$$

Thus, the final loss function to learn the Φ -space is given by

$$\mathbb{L}_{\Phi} = \lambda_1 \mathbb{L}_{latent} + \lambda_2 \mathbb{L}_{recons} \quad (3)$$

where, the weights for the losses λ_1 and λ_2 are computed based on the retrieval accuracy on a validation set.

3.2. Style Module

In s-SBIR, the user has the flexibility to specify their preferred styles in the form of keywords (e.g., *red* in color, *rectangular* in shape etc.). For this purpose, two style encoders are required, one (E_{style}^a) for encoding the style from the keywords (given with the query sketch), and the other (E_{style}^I) for encoding the style from the images.

Encoding Style from Keywords: While training, we do not have access to the complete query-pair, i.e. {sketch, user-defined style-keywords} along with the ground-truth retrieved images. Hence, we treat the image-style annotations \mathbf{a}_i^I as the substitute for style-requirements, for which \mathbf{x}_i^I is the perfectly matched sample.

In many existing works [24][49], the binary style vector \mathbf{a}_i^I is used directly for computation. Since binary representations do not capture the relations within different

styles (eg., *red* is closer to *pink* compared to *blue*), it restricts the query style to belong to one of the styles used in training. Here, we again use the GloVe-embeddings of the style-keywords as its representation for better generalization. Given \mathbf{a}_i^I , for each style present, we obtain its GloVe-embedding and then concatenate them to obtain the complete style-vector \mathbf{p}_i . In case the sample is not annotated for all types of styles (or, the user provides only *color* requirement and no other style information), we use a globally (across the dataset) computed initialization for that missing style. In this work, we use the mean of the GloVe-features of all the possible values of a particular style in the training data for initialization. Finally, \mathbf{p}_i is passed through the encoder E_{style}^a to obtain the final style-annotation representation of \mathbf{x}_i^I .

Encoding Style from Images: Since the goal is to retrieve images having the desired styles given by the keywords, we want to match the style encoding from the keywords to the style extracted from the images. Thus, we aim to extract the style information \mathbf{z}_i^I from the image \mathbf{I}_i such that it represents the ground-truth style annotation \mathbf{a}_i^I of that image.

Computation of \mathbf{z}_i^I is based on the intuition that image-specific styles are embedded in the initial layers of the classification network [11][16] (VGG-19 in our case) and only the class-specific high level information is present in the final layers. Thus, we capture the style information from the activation maps of the lower or middle-layers [11] of FGI . For an input image \mathbf{I}_i , let the activation maps for the l^{th} layer be denoted as $\mathbf{V}_l(\mathbf{I}_i) \in \mathbb{R}^{d_l^V \times \eta_l}$, where $d_l^V = M_l \times N_l$ and M_l, N_l are the height and width of the activation map respectively. η_l is the number of convolution filters in the l^{th} layer. The image-specific style in the l^{th} layer is captured as the Gram matrix $\mathbf{G}_l(\mathbf{I}_i) \in \mathbb{R}^{\eta_l \times \eta_l}$ formed by the activations of the layer as [11],

$$\mathbf{G}_l(\mathbf{I}_i) = \frac{1}{d_l^V \times \eta_l} \mathbf{V}_l^T(\mathbf{I}_i) \cdot \mathbf{V}_l(\mathbf{I}_i) \quad (4)$$

We obtain \mathbf{z}_i^I by concatenating this representation $\mathbf{G}_l(\mathbf{I}_i)$ (after vectorization) obtained from several layers $l \in \mathcal{L} = \{1, 2, \dots, L\}$. The value of L can be set experimentally. In this work, we use $L = 2$ (VGG-layers ‘conv1_1’, ‘conv2_1’), hence the concatenated \mathbf{G}_l -vector is of 20480-d, which is then reduced to 1024-d after PCA. This representation is then applied as input to the encoder E_{style}^I , such that the output encoding matches with $E_{style}^a(\mathbf{p}_i)$ obtained from \mathbf{a}_i^I . Thus our style-space loss is formulated as

$$\mathbb{L}_{style} = \|E_{style}^a(\mathbf{p}_i) - E_{style}^I(\mathbf{z}_i^I)\|_2^2 + \lambda_{style}^a \|\theta_{style}^a\|^2 + \lambda_{style}^I \|\theta_{style}^I\|^2 \quad (5)$$

Here λ_{style}^a and λ_{style}^I are the two hyper-parameters which are set empirically and θ_{style}^a and θ_{style}^I are the learnable parameters for the style encoders.

3.3. Mixer module

Given a sketch with desired keywords, the goal of s-SBIR is to retrieve the relevant images from the correct category with the desired styles. To achieve this, we propose the mixer network N_{mixer} such that it transforms the concatenated input of sketch category information $E_S(\mathbf{x}_j^S)$ and style obtained from the keywords $E_{style}^a(\mathbf{p}_i)$ to form a composite representation $\mathbf{s}_j^{c,a}$ in a latent space Ψ . Similarly, the concatenated vector of image-category $E_I(\mathbf{x}_i^I)$ and extracted image-style $E_{style}^I(\mathbf{z}_i^I)$ is transformed using the same network N_{mixer} , such that this composite image-representation $\mathbf{i}_i^{c,a}$ can be compared directly with $\mathbf{s}_j^{c,a}$ for retrieval. Here, with slight abuse of notations, the superscript c and a denotes the category and style of the input.

Though the images retrieved in this Ψ -space should have the correct category and user-specified styles, still the importance of the two components are different. An user would prefer an image from the correct category (even with different styles) over images from an incorrect category (even with the same styles). Keeping this in mind, the mixer network N_{mixer} is designed using the following criteria

$$d(\mathbf{s}_j^{c,a}, \mathbf{i}_*^{c^+,a^+}) < d(\mathbf{s}_j^{c,a}, \mathbf{i}_*^{c^+,a^-}) < d(\mathbf{s}_j^{c,a}, \mathbf{i}_*^{c^-,a^+}) < d(\mathbf{s}_j^{c,a}, \mathbf{i}_*^{c^-,a^-}) \quad (6)$$

Here, $\mathbf{i}_*^{c^+,a^+}$ represents the images (in the Ψ -space) belonging to the same category c as the input sketch and having the same style a . Similarly, $\mathbf{i}_*^{c^-,a^-}$ represents all the images from some class other than c and having a different style than a and so on. $d(\mathbf{m}, \mathbf{n})$ represents the Euclidean distance between the vectors \mathbf{m} and \mathbf{n} .

To impose the above condition (6), we construct a quintuplet set [14], $\mathcal{Q} = \{\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^+,a^+}, \mathbf{i}_j^{c^+,a^-}, \mathbf{i}_j^{c^-,a^+}, \mathbf{i}_j^{c^-,a^-}\}_{j=1}^N$, based on the image category and styles. N is the total number of quintuplet-instances selected. We formulate the loss function to learn the parameters of N_{mixer} (θ_{mixer}) subject to the condition (6)

$$\mathbb{L}_{mixer} = \min \sum_{j \in \mathcal{Q}} \alpha_j + \beta_j + \gamma_j + \lambda \|\theta_{mixer}\|_2^2, \quad s.t.$$

$$\max(0, m_1 + d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^+,a^+}) - d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^+,a^-})) < \alpha_j,$$

$$\max(0, m_2 + d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^+,a^-}) - d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^-,a^+})) < \beta_j,$$

$$\max(0, m_3 + d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^-,a^+}) - d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^-,a^-})) < \gamma_j$$

where $\alpha_j, \beta_j, \gamma_j \geq 0$ are the slack variables and m_1, m_2 and m_3 are the margins set experimentally. λ is a regularization parameter.

3.4. Proposed Normalized Composite MAP

The standard metric used for SBIR evaluation is Mean Average Precision (MAP) [43][25], which considers only

the category of the input sketch and the database images. Since the styles of the retrieved examples are not considered in this evaluation, this cannot be directly used for s-SBIR approaches. Even for single modality, style-guided retrieval approaches are demonstrated through qualitative results only [24][49]. Here, we propose a generalization of the standard MAP, termed normalized composite-MAP (or ncMAP) for evaluation of such approaches which can be used for both single and cross-modal applications. Note that the proposed ncMAP reduces to MAP if we consider only the categories.

We propose to assign two separate scores for each retrieved image, one based on its category and the other based on its style. The score for the k^{th} retrieved image \mathbf{i}_k against a query \mathbf{s}_q is given by

$$\begin{aligned} score_{cat}(\mathbf{s}_q, \mathbf{i}_k) &= 1, \text{ if category}(\mathbf{s}_q) = \text{category}(\mathbf{i}_k) \\ &= 0, \text{ otherwise} \end{aligned}$$

Let us assume the style given by the keywords is represented as a binary vector $\mathbf{a}_q^S \in \mathbb{Z}^{d_a}$, where d_a is the total number of styles. An entry of 1 indicates that style is desired, otherwise, it is 0. Given that the style annotations for the k^{th} image is \mathbf{a}_k^I , we assign another score, $score_{style}(\mathbf{a}_q^S, \mathbf{a}_k^I)$,

$$score_{style}(\mathbf{a}_q^S, \mathbf{a}_k^I) = \text{cosine_similarity}(\mathbf{a}_q^S, \mathbf{a}_k^I)$$

Hence for a given sketch and style-keywords as query, for each k^{th} retrieved image, we assign the composite score as

$$\begin{aligned} c\delta_w(\mathbf{s}_q, \mathbf{a}_q^S, \mathbf{i}_k, \mathbf{a}_k^I) &= w * score_{cat}(\mathbf{s}_q, \mathbf{i}_k) \\ &+ (1 - w) * score_{cat}(\mathbf{s}_q, \mathbf{i}_k) * score_{style}(\mathbf{a}_q^S, \mathbf{a}_k^I) \end{aligned}$$

where w is the weighting factor which controls the importance of category-match over styles. $c\delta_w$ reaches maximum value to 1, if both the category and style of the retrieved image matches with that of the query. It reduces to zero when category is not matched, even though the style-requirement is matched. Hence, the composite precision at k^{th} retrieved image is assigned as $cP_w = \frac{1}{k} \sum_{r=1}^k c\delta_w(\mathbf{s}_q, \mathbf{a}_q^S, \mathbf{i}_r, \mathbf{a}_r^I)$. We compute the composite average precision (cAP_w) over top- K retrieved images as

$$cAP_w@K = \frac{1}{TP} \sum_{r=1}^K [score_{cat}(\mathbf{s}_q, \mathbf{i}_r) * cP_w(\mathbf{s}_q, \mathbf{a}_q^S, \mathbf{i}_r, \mathbf{a}_r^I)]$$

TP is the total correct retrieved elements (category-wise) in top- K . Finally, the composite-MAP ($cMAP_w$) is measured over the entire query set \mathcal{Q} as,

$$cMAP_w@K = \frac{1}{|\mathcal{Q}|} \sum_{q=1}^{|\mathcal{Q}|} cAP_w(\mathbf{s}_q, \mathbf{a}_q^S, K)$$

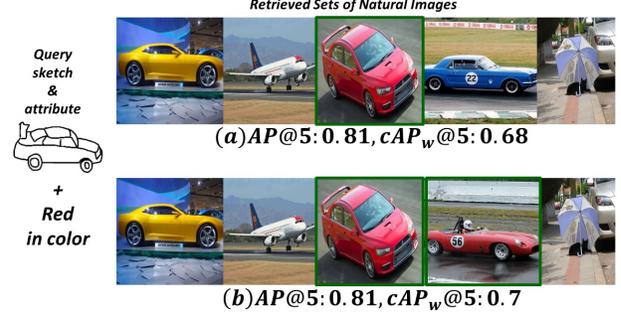


Figure 3. Illustration of proposed evaluation metric: Given a sketch query of *car* and preferred color as *red*, the cAP_w score for the second retrieved image set is more than the first set, though standard AP, which considers only the category, is same for both.

We further normalize this $cMAP$ with $cMAP^{ideal}$, which is evaluated using the ideal retrieved list based on the ground-truth categories and styles. Thus, $ncMAP_w@K = \frac{cMAP_w@K}{cMAP_w^{ideal}@K}$. The highest value of this proposed metric is 1, irrespective of the dataset. Figure 3 illustrates how the proposed ncMAP takes into consideration both the category and style information. All the results reported in terms of ncMAP in this paper use $w = 0.8$.

3.5. Testing

Given a query sketch with the desired style-keywords in the form of \mathbf{a}_q^S and a set of images, we obtain the respective feature representations from FG_S and FG_I -subnetworks as \mathbf{s}_q and $\mathcal{I} = \{\mathbf{i}_1, \dots, \mathbf{i}_n\}$, respectively. We further obtain the Φ -space representations of the query sketch and image set. The representation of the query style is computed as $E_{style}^a(\mathbf{p}_q)$ and of the image styles as $E_{style}^I(\mathbf{z}_1^I), \dots, E_{style}^I(\mathbf{z}_n^I)$. In this case, \mathbf{p}_q (from \mathbf{a}_q^S) and $\mathbf{z}_i^I, i = 1, \dots, n$ are obtained as described in Section 3.2. Then, we find the compact category-style representations of the samples in Ψ -space using the mixer network as $N_{mixer}(E_S(\mathbf{s}_q), E_{style}^a(\mathbf{p}_q))$ and $N_{mixer}(E_I(\mathbf{i}_k), E_{style}^I(\mathbf{z}_k^I))$, for $k \in \{1, \dots, n\}$. Finally, we use the Euclidean distance between the query and image samples in the Ψ -space to get the retrieved images.

3.6. Difference with [6]

The work in [6] studies a similar problem as s-SBIR, though there are significant differences. The style component in the style query accompanying the sketch in [6] is provided as an image, the aesthetic component of which is the required style. In contrast, s-SBIR has the flexibility to provide the style-requirement (single / multiple) in terms of simple keywords, which eliminates the need to find an image reflecting all the requirements.

Styles	Values	Total
color	<i>black, blue, brown, ...</i>	11
texture	<i>spots, stripes, ...</i>	3
shape	<i>round, columnar, ...</i>	4
material	<i>metal, wooden, ...</i>	3
structure	<i>bipedal, quaduped, ...</i>	4

Table 1. Few examples of image styles in the datasets used.

4. Experimental Evaluation

For training and testing s-SBIR approaches, we need the ground-truth styles of the images. Because of the unavailability of such large-scale datasets, we propose a new split for two standard SBIR datasets, namely Sketchy and TU-Berlin, whose object categories are a subset of ImageNet [31]. A part of the ImageNet image instances are annotated with styles (color, shape, material etc.) [24] and we select those categories from both datasets for which instance-level annotations are available. We now give a brief description of the modified datasets used for this work. **The m-Sketchy Database:** The original dataset [34] is a collection of around 75,000 sketches and 12,500 images from 125 object categories. For our experiment, we use the Sketchy extension [25] which has 73,000 images. To evaluate s-SBIR approach, we construct a modified Sketchy or m-Sketchy dataset, which has 31,378 sketches and 31,900 images from 53 categories (~ 600 samples / class).

The m-TU-Berlin Dataset [8]: The original dataset contains sketch data from 250 object categories and for each category, 80 sketch samples are available. We use additional 204,489 natural images, provided by [48] as the image data [25][43]. The modified m-TU-Berlin contains 7600 sketches and 8587 images from 95 classes (80 sketches / class and ~ 45 -150 images / class).

The ground-truth image styles [24] are 25-d vectors with values representing negative (-1), unsure (0), positive ($+1$) and missing (2) styles/attributes. We consider negative, unsure, missing as negative (0) in this work. Table 1 demonstrates sample styles and their values.

We use 10% of sketch-image data in m-Sketchy and m-TU-Berlin for testing and the rest for training. We construct the test set for s-SBIR by augmenting the query sketches with possible styles obtained from the images belonging to the same category as the sketch in the training data. Figure 4 depicts a pictorial description of the test set construction process. Once constructed, this augmented test set is fixed and all the evaluations are performed on this new test-set for fairness. We implement our framework using TensorFlow [1]. The detailed network architecture is summarized in Table 2. For training, Adam optimizer is used with learning rate $1e-3$, momentum 0.9 and batch size 32. All the hyper-parameters are set based on the accuracy on the validation set, which is constructed as 10% of the training data.

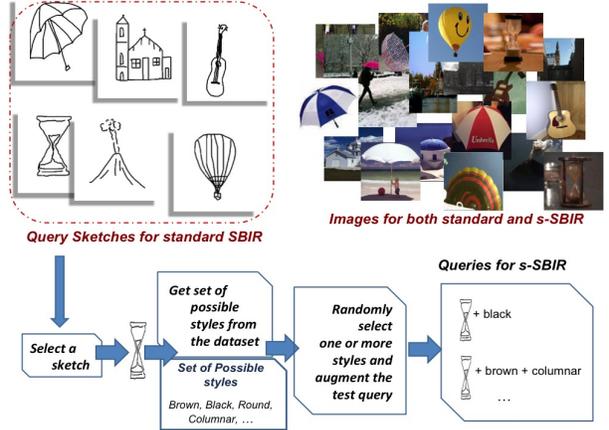


Figure 4. Flowchart depicting the test set construction for s-SBIR.

Network	Description	I/p, O/p (O/p-d)
FG_I	VGG-19 w/o final softmax	$\mathbf{I}_i, \mathbf{x}_i^I$ (4096-d)
FG_S	VGG-19 w/o final softmax	$\mathbf{S}_j, \mathbf{x}_j^S$ (4096-d)
E_m	$fc:2, N_1:1024, N_2:200, \text{ReLU}$	$\mathbf{x}_i^m, E_m(\mathbf{x}_i^m)$ (200-d)
D_m	$fc:2, N_1:1024, N_2:4096, \text{ReLU}$	$E_m(\mathbf{x}_i^m), D_m(E_m(\mathbf{x}_i^m))$ (4096-d)
E_{style}^I	$fc:2, N_1:500, N_2:200, \text{ReLU}$	$\mathbf{z}_i^I, E_{style}^I(\mathbf{z}_i^I)$ (200-d)
E_{style}^a	$fc:2, N_1:200, N_2:200, \text{ReLU}$	$\mathbf{p}_i, E_{style}^a(\mathbf{p}_i)$ (200-d)
N_{mixer}	$fc:1, N_1:200, \text{ReLU}$	$\{E_I(\mathbf{x}_i^I), E_{style}^I(\mathbf{z}_i^I)\},$ $\mathbf{i}_i^{c;p}$ (200-d) or, $\{E_S(\mathbf{x}_j^S), E_{style}^a(\mathbf{p}_i)\},$ $\mathbf{s}_j^{c;p}$ (200-d)

Table 2. Architecture of the proposed s-SBIR framework. For each sub-network, the description is reported in the format {type and number of layers, number of nodes in i^{th} layer (N_i), activation function}. The final column describes the input (I/p), output (O/p), and the dimension of output (O/p-d) for each sub-network. $m \in \{I, S\}$ for image and sketch respectively.

Baselines: Since, there exists no baseline to directly compare with s-SBIR, we develop different baselines to address the same problem and use them to compare and analyze the proposed framework.

Baseline 1: We choose the first baseline to be the proposed category module (CM), which can be used for standard SBIR to show that as expected, the retrieved images, even though they are of the correct category, are arranged in a random manner with respect to style.

Baseline 2: For this baseline, CS_{fused} , we perform s-SBIR based on score level fusion of category-based and style-based similarities of the style-augmented sketch query with the database images. Given a query sketch s_q and a set of natural images $\mathcal{I} = \{\mathbf{i}_1, \dots, \mathbf{i}_n\}$, we compute their category-similarity scores $\text{sim}_{cat} \in \mathbb{R}^n$ based on their respective Φ -space representations. In addition, given the query style-keyword \mathbf{a}_q^S , we obtain the style-similarity scores $\text{sim}_{style} \in \mathbb{R}^n$, using $E_{style}^a(\mathbf{p}_q)$ and $E_{style}^I(\mathbf{z}_k^I), k = 1, \dots, n$. Finally, the fused score is computed for each im-

Model	Datasets			
	m-Sketchy		m-TU-Berlin	
	MAP	ncMAP _w	MAP	ncMAP _w
B1 - CM	0.7325	0.7569	0.6647	0.6722
B2 - CS_{fused}	0.6113	0.7084	0.6679	0.6919
B3 - $CS_{triplet}$	0.3758	0.4943	0.4325	0.5212
B4 - $CS_{Kron-fusion}$	0.7182	0.7320	0.6111	0.6598
B5 - $CS_{fbilinear-fusion}$	0.7377	0.7818	0.6599	0.7150
Proposed	0.7481	0.7925	0.6806	0.7400

Table 3. Evaluation of s-SBIR on m-Sketchy and m-TU-Berlin.

age in \mathcal{I} using a convex combination of both scores as, $sim_{fused} = \alpha sim_{cat} + (1 - \alpha) sim_{style}$. This is used to retrieve the relevant images for the query. We have taken $\alpha = 0.8$ as the fusion weight.

Baseline 3: For this baseline $CS_{triplet}$, we employ a triplet loss on the combination of category and style. The required triplet loss is formulated as follows

$$\mathbb{L}_{triplet} = \min \sum_{j \in \mathcal{Q}} \alpha_j + \lambda \|\theta_{mixer}\|_2^2,$$

$$\text{s.t. } \max(0, m_1 + d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^+,a^+}) - d(\mathbf{s}_j^{c,a}, \mathbf{i}_j^{c^*,a^*})) < \alpha_j$$

$\mathbf{i}_j^{c^+,a^+}$ essentially considers any combination of category and style (except $\mathbf{i}_j^{c^+,a^+}$) as a negative sample against $\mathbf{s}_j^{c,a}$. Here we give equal weightage to the category and style.

Baseline 4 and 5: For the last two baselines, we use variants of how to combine the content and style information using the mixer module. As described in Section 3.3, the category-style mixer network essentially fuses the category and style information into a single composite representation in Ψ -space. The proposed framework uses concatenation-based fusion. In Baseline 4, we employ the Kronecker-product-based fusion as used in sketch-image hashing [35] and this is denoted as $CS_{Kron-fusion}$. For Baseline 5, we fuse the content and style using the popular factorized bi-linear pooling widely used in visual question-answering [46] etc. This is denoted as $CS_{fbilinear-fusion}$.

The performance of the proposed framework along with the different baselines in terms of MAP and ncMAP_w is reported in Table 3. We make the following observations - (1) Except **B3**, the other baselines gives primary importance to the category, so their MAP is better as compared to **B3**; (2) **B3** gives equal importance to category and style which results in poor MAP, which also results in poor ncMAP_w; (3) The relative increase from MAP to ncMAP_w is much less for **B1** signifying that even though the retrieved categories are correct, the styles are random as expected.

Qualitative results: Figure 5(b) shows the top-10 retrieved images for few sketch queries with single desired style. We observe that the desired style is pushed to the top results while still maintaining the correct category.

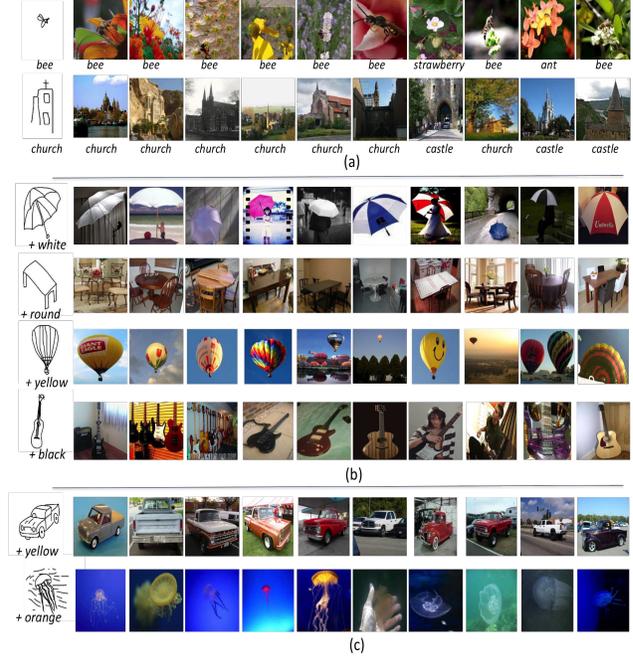


Figure 5. Qualitative results on m-Sketchy dataset. (a) Top-10 retrieved images using the proposed category module for SBIR. (b) and (c) shows the top-10 retrieved results for sketch query and a single style criteria, when the style is (b) present; (c) not present in the search set. Figure best viewed in color.

We also experimented with query styles which are not present in the search set (In Figure 5(c)). For this experiment, for a given query and desired style, e.g. truck images with desired color *yellow*, we remove all the truck images annotated as *yellow* in color from the search set and then perform retrieval. The results show that the top-retrieved images resemble perceptually similar colors (*orange*, *red*) to *yellow*, compared to very different ones (*black*) present in the dataset. This result corroborates our idea that using GloVe-feature representation of individual styles instead of the binary representation [24] provides a semantically meaningful style-space. A similar experiment has also been performed for query $\{jellyfish-sketch, orange\}$.

Figure 6 shows some qualitative retrieval results when the input sketch is augmented with multiple (in this case two) styles. Even though there are multiple styles, the proposed framework is able to combine them with the category information in an effective manner to retrieve images of the correct category with a combination of the desired styles. In our experiments, we have used either single or two-styles, but the proposed framework can seamlessly be extended for more than two styles as well.

4.1. Analysis

In this section, we report the analysis on the m-Sketchy dataset, unless mentioned explicitly.

Standard Category-based SBIR: Here, we compare the



Figure 6. Top-5 results for s-SBIR using the proposed framework on m-Sketchy data when multiple desired styles are given as query.

category module of the proposed approach with standard SBIR approaches, which consider only the category of the query sketch and the retrieved images. Note that the proposed category module is a generic one and can potentially be replaced with a better module, while still maintaining the style and mixer modules intact. We compare the proposed category module with state-of-the-art cross-modal hashing [25][3] approaches on the full Sketchy dataset using the same split [25] as used by the other algorithms.

	Method	feature dim.	MAP
previous SBIR methods	GF-HOG [13]	3500	0.157
	SHELO [32]	1296	0.161
	LKS [33]	1350	0.190
	Siamese-CNN [29]	64	0.481
	SaN [45]	512	0.208
	GN Triplet [34]	1024	0.529
	3D Shape [40]	64	0.084
	Siamese-AlexNet	4096	0.518
	Triplet-AlexNet	4096	0.573
Cross-modal Hashing methods (128-bits)	CMFH [7]		0.190
	CMSSH [4]		0.211
	SCM-Seq [47]		0.671
	SCM-Orth [47]	4096	0.616
	CVH [20]		0.624
	SePH [23]		0.640
	DCMH [17]		0.656
	DSH [25](32-bits)		0.653
	DSH [25](128-bits)		0.783
Cross-modal feature learning (continuous-value)	CCA [38]		0.705
	XQDA [22]		0.550
	PLSR [41]	4096	0.462
	CVFL [42]		0.675
	Proposed	200	0.7968

Table 4. Comparison of our Category Module with state-of-the-art cross-modal hashing approaches on Sketchy for SBIR.

The results are reported in Table 4. We observe that on Sketchy data, the proposed model outperforms the state-of-the-art DSH [25] with much smaller feature dimension. This is surprising given the simple architecture of the proposed category module (*CM*), since the main focus of this work is the integration of style information. Note that our *CM* does not use any privileged information about the data, such as edge maps [25].

Latent space variations: We formed the latent space for

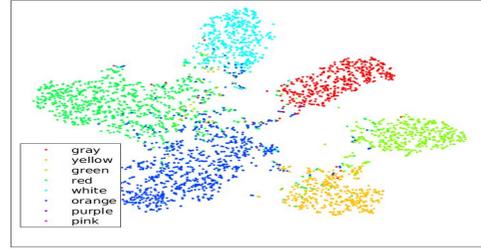


Figure 7. t-SNE plot of extracted color features from images, where each image is annotated with only one color.

SBIR using the category-name embeddings extracted from pre-trained GloVe [28] model. Effectively, this makes the Φ -space similar to the label-space of training data which is used in few recent SBIR work [25]. We perform an experiment using these two variations of label/category encoding to form the Φ -space and observe their effects on the category-based SBIR in the latent Φ -space. We obtain a MAP of 0.7076 and 0.7325 using one-hot encoding of labels and Glove-embeddings respectively. Furthermore, using Glove-embeddings as latent-space results in a semantically meaningful retrieved set, where even the incorrectly retrieved objects are semantically related to the query object, even though they are not exactly same. This is reflected in the visual results shown in Figure 5(a). Against a query of sketch *bee* or *church*, the model has wrongly retrieved images of *ant* or *castle*, which have strong visual (shape/structure-wise) as well as semantic similarity with the query objects. However, using one-hot encoding of labels may lack this inter-category semantic relationship. **Effective style-space construction:** Here, we analyse the effectiveness of the lower and middle-layer VGG-features of an image for style extraction. For ease of visualization and understanding, we choose those images which are annotated with single value of a particular style, color. We observe from the t-SNE [27] plot of the extracted image style-features in Figure 7 that the features form nice clusters in the style-space (or color-space), which justifies the usefulness of the style extraction approach.

5. Conclusion

Here, we have introduced the problem of s-SBIR, where the user has the flexibility of specifying any desired style of the retrieved images. We have proposed an end-to-end deep framework, which uses a category module, style module and mixer module to appropriately disentangle and mix the category and style information for s-SBIR. We have also proposed a composite metric to evaluate s-SBIR approaches. Extensive experiments and analysis on subsets of two widely-used sketch-image datasets show the effectiveness of the proposed framework.

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.
- [2] A. D. Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 19(2):121–132, 1997.
- [3] K. Bozas and E. Izquierdo. Large scale sketch based image retrieval using patch hashing. In *Proceedings of International Symposium on Visual Computing*, 2012.
- [4] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2010.
- [5] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: interpretable representation learning by information maximizing generative adversarial nets. *arXiv:1606.03657v1*, 2016.
- [6] J. Collomosse, T. Bui, M. Wilber, C. Fang, and H. Jin. Sketching with style: visual search with sketches and aesthetic context. In *Proceedings of International Conference on Computer Vision*, 2017.
- [7] G. Ding, Y. Guo, and J. Zhou. Collective matrix factorization hashing for multimodal data. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2014.
- [8] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM Transactions on Graphics*, 31(4):44:1–44:10, 2012.
- [9] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: benchmark and bag-of-features descriptors. *IEEE Transactions in Visual Computation Graph*, 17(11):1624–1636, 2011.
- [10] T. Furuya and R. Ohbuchi. Hashing cross-modal manifold for scalable sketch-based 3d-model retrieval. In *Proceedings of International Conference on 3D Vision*, 2014.
- [11] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of IEEE Computer Vision Pattern Recognition*, 2016.
- [12] N. Hadad, L. Wolf, and M. Shahar. A two-step disentanglement method. In *Proceedings of Computer Vision Pattern Recognition*, 2018.
- [13] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *Proceedings of IEEE International Conference in Image Processing*, 2010.
- [14] C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2016.
- [15] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, and W. Fan. Sketch-based image retrieval with deep visual semantic descriptor. *Pattern Recognition*, 76:537–548, 2018.
- [16] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of European Conference in Computer Vision*, 2018.
- [17] Q. Y. Jiang and J. W. Li. Deep cross-modal hashing. In *Proceedings of IEEE Computer Vision Pattern Recognition*, 2017.
- [18] T. Kato, T. Kurita, N. Otsu, and K. Hirata. A sketch retrieval method for full color image database-query by visual example. In *Proceedings of International Conference on Pattern Recognition*, 1992.
- [19] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of Computer Vision Pattern Recognition*, 2017.
- [20] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2011.
- [21] Y. Li, T. M. Hospedales, Y. Z. Song, and S. Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *Proceedings of British Machine Vision Conference*, 2014.
- [22] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2015.
- [23] Z. Lin, G. Ding, M. Hu, and J. Wang. Semantic preserving hashing for cross-view retrieval. In *Proceedings of IEEE Computer Vision Pattern Recognition*, 2015.
- [24] H. Liu, R. Wang, S. Shan, and X. Chen. Learning multifunctional binary codes for both category and attribute oriented retrieval tasks. In *Proceedings of the IEEE Computer Vision Pattern Recognition*, 2017.
- [25] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE Computer Vision Pattern Recognition*, 2017.
- [26] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE Computer Vision Pattern Recognition*, 1999.
- [27] L. V. D. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(8):2579–2605, 2008.
- [28] R. J. Pennington and C. Manning. Glove: global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2014.
- [29] Y. Qi, Y. Z. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *Proceedings of the IEEE International Conference in Image Processing*, 2016.
- [30] Y. Z. Qi, Y. amd Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo. Making better use of edges via perceptual grouping. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2015.
- [31] 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. *arXiv:1409.0575[cs]*, 2014.
- [32] J. Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *Proceedings of the IEEE International Conference in Image Processing*, 2014.

- [33] J. M. Saavedra and J. M. Barrios. Sketch based image retrieval using learned keyshapes (lks). In *Proceedings of British Machine Vision Conference*, 2015.
- [34] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics*, 35(4):1–12, 2016.
- [35] Y. Shen, L. Liu, F. Shen, and L. Shao. Zero-shot sketch-image hashing. In *Proceedings of Computer Vision Pattern Recognition*, 2018.
- [36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556[cs]*, 2014.
- [37] 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 *Proceedings of International Conference on Computer Vision*, 2017.
- [38] B. Thompson. Canonical correlation analysis. *Encyclopedia of Statistics in Behavioral Science*, 2005.
- [39] N. Vo, L. Jiang, C. Sun, K. Murphy, L. J. Li, L. Fei-Fei, and J. Hays. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of Computer Vision Pattern Recognition*, 2019.
- [40] F. Wang, L. Kang, and Y. Li. Sketch-based 3d shape retrieval using convolutional neural network. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2015.
- [41] H. Wold. Partial least squares. *Encyclopedia of Statistical Sciences*, 1985.
- [42] W. Xie, Y. Peng, and J. Xiao. Cross-view feature learning for scalable social image analysis. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2014.
- [43] 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 Transactions in Image Processing*, 27(9):4410–4421, 2018.
- [44] Q. Yu, F. Liu, Y. Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy. Sketch me that shoe. In *Proceedings of Computer Vision Pattern Recognition*, 2016.
- [45] Q. Yu, Y. Yang, F. Liu, Y. Z. Song, T. Xiang, and H. Hospedales. Sketch-a-net: a deep neural network that beats humans. *International Journal of Computer Vision*, 112(3):411–425, 2017.
- [46] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of International Conference on Computer Vision*, 2017.
- [47] D. Zhang and W. J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2014.
- [48] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. Sketchnet: sketch classification with web images. In *Proceedings of IEEE Computer Vision Pattern Recognition*, 2016.
- [49] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Computer Vision Pattern Recognition*, 2017.
- [50] W. Zhou, H. Li, and Q. Tian. Recent advance in content-based image retrieval: a literature survey. *arXiv preprint arXiv:1706.06064*, 2017.