# StyleGuide: Zero-Shot Sketch-based Image Retrieval Using Style-Guided Image Generation

Titir Dutta, Anurag Singh, Soma Biswas, *Senior Member, IEEE*

***Abstract*—The goal of zero-shot sketch-based image retrieval is to retrieve relevant images from a search set against a hand-drawn sketch query, which belongs to a class, previously unseen by the model. The knowledge gap between such unseen and seen classes along with the domain-gap between the query and search-set makes the problem extremely challenging. In this work, we address this problem by proposing a novel retrieval methodology, *StyleGuide* using style-guided *fake*-image generation. In addition, we further study the scenario of generalized zero-shot sketch-based image retrieval, where the search set contains images from both seen and unseen categories. Specifically, we propose an unseen class sample detection approach based on pre-computed prototypes to construct a refined search set for such experimental settings. Thus, the query sketch needs to be compared only to those image data which are more likely to belong to the unseen classes, resulting in improved retrieval performance. 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 for ZS-SBIR and gives significant improvements over the state-of-the-art for generalized zero-shot sketch-based image retrieval.**

***Index Terms*—sketch-based retrieval, generalized ZS-SBIR, content-style decomposition, novelty detection**

## I. Introduction

SKETCH-based Image Retrieval (SBIR) is a vey relevant problem in today's era of smart gadgets and touch-screen devices with potential applications in the field of e-commerce, forensics, etc. It allows an user the flexibility to search an image database with a roughly-drawn sketch, instead of writing a text description (text-based image retrieval) or obtaining a similar image (image-to-image retrieval). The problem of SBIR has been considered from two different perspectives in literature. Retrieval of image instance from a database matching the exact details (shape / pose) of the hand-drawn sketch query is referred to as fine-grained SBIR [1][2]. On the other hand, category-based SBIR [3][4] considers retrieving images of same category as in the sketch, irrespective of their shape / pose. However, during retrieval, the sketch query and the database images are always considered to belong to one of the training categories, which ensures that the trained model has adequate knowledge of the sketch-image mapping.

In real-life applications, objects of new categories are continuously being added to the database and thus the above assumption requires the model to be re-trained every time. To address such an issue, zero-shot sketch-based image retrieval (ZS-SBIR) [5][6] is gaining increasing attention, where the query sketches and database images belong to *novel* categories, which are not seen during training. A degraded retrieval performance of standard SBIR methods has been observed in [6] for such protocol. A generative-model based retrieval [6] along with fusion-network based latent-space learning [5], semantic-aware cycle-consistent network [7] have been proposed in recent literature to address ZS-SBIR.

We address ZS-SBIR by separating both sketches and images into their domain-independent *content* and domain and data specific variations or *styles*. The *content* lies in a latent-space shared by both domains [4][3] and follow a meaningful semantic order. We propose to obtain the initial list of retrieved images by matching the *content*-information extracted from query sketch and database images in this latent space. However, motivated by the better recognition performance in the image-space, compared to the semantic-space, as in [8] (for the task of zero-shot learning), we further propose to refine the latent-space rank-list based on the matching in image-space. Specifically, given the top-K images based on *content*-matching in the latent-space, we fuse the sketch query *content* with the specific *style*s of each of these images to generate K-fake images, which are finally used to re-rank the initial retrieved list. We refer to the proposed style-guided generation-based retrieval network as *StyleGuide*.

An extension to the ZS-SBIR protocol is discussed in [7], where the search-set contains images from both seen and unseen categories and is referred to as generalized ZS-SBIR. Intuitively, the retrieval performance should decrease in such scenario, since the presence of seen class images in the search set poses a higher degree of confusion for the algorithm. To mitigate this, we propose a novel approach for detecting the unseen class images in the search set prior to retrieval. Since the query sketch belongs to an unseen class, the images, predicted to belong to the seen class set by the proposed algorithm, can be removed from the search set. This allows the query to be searched against a smaller subset of the database, resulting in faster retrieval and improved retrieval accuracy. The key-contributions of this work are summarized below.

- 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 ZS-SBIR, by separating the original data representations into a semantic-aware domain-invariant content and domain and data specific variations/style.
- We propose a novel unseen class detection mechanism to reduce the search set for generalized ZS-SBIR for improved retrieval performance.
- Experiments on two large-scale sketch-image datasets, Sketchy extended [9] and TU-Berlin [10], and extensive analysis with different variants of the proposed approach

show the effectiveness of the proposed framework.

This is an extended version of our previous work [11], which proposes a content-style disentanglement based retrieval framework for ZS-SBIR and generalized ZS-SBIR. The rest of the paper is organized as follows. Section II gives a brief description of the related work. Our proposed approach for ZS-SBIR is discussed in Section III, and that for generalized ZS-SBIR in Section IV. This is followed by extensive experiments (Section V) and analysis (Section VI) and conclusion in Section VII.

## II. Related work

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

**Sketch-based Image retrieval (SBIR):** The domain-gap between sketch and image representations is the main challenge addressed in SBIR. Early methods used specially-designed hand-crafted features [12][13] for both sketches and the edge-maps extracted from images to account for the domain-difference. With the advances in deep learning, siamese networks [14], triplet-loss [9] or contrastive-loss [15] based end-to-end learning models, as well as their combinations [16] have been proposed for the same. [3] proposes a heterogeneous network, which exploits deep-features of sketches, images and the edge-maps to retrieve images for a hashing-based SBIR protocol. A generative model is proposed to generate images from sketches towards the goal of hashing-based retrieval in [17]. A non-deep shared-space learning method exploiting curriculum learning is proposed in [4]. [18] addresses a variant of SBIR, where the preferred *aesthetic style* of the retrieved images is also specified with the query.

**Zero-shot Sketch-based Image Retrieval (ZS-SBIR):** The experimental protocol for SBIR is generalized to ZS-SBIR [5][6][7], where the sketch query and database images are from categories other than the training categories. In [5], a sketch-image feature fusion-based end-to-end model is proposed, which has a very high memory requirement [7]. [6] propose a sketch-to-image feature generation and voting-based image retrieval at the image feature-space; but restricts the model to be learned with paired data only. A semantically-aligned latent-space learning is proposed in [7], which involves learning two GANs. A new large-scale sketch-image dataset is proposed in [19], and a triplet-loss based network is presented to address ZS-SBIR.

**Zero-shot learning (ZSL):** A related research area is ZSL, which addresses the problem of classifying images from unknown categories. A latent-space learning approach [20][21] is quite popular for the same, which transforms the image features and class-semantic information (attributes) to a common-space for matching. In contrast, synthsizing image features for unseen classes, using generative models [8][22][23] and training a classifier using the same have obtained state-of-the-art performance for ZSL.

**Novelty Detection:** Novelty detection [24] or out-of-distribution sample detection [25] is a popular research topic in computer vision. Such techniques have been used to address the problem of generalized ZSL in [26][27][28]. While [26][27] exploits novelty detection to decide on the choice of classifier for the test sample, [28] uses an auto-encoder (AE) based network which demonstrates a higher reconstruction error for novel class samples.

**Content-Style disentanglement:** The successful use of content-style disentanglement [29][30] for various applications of computer vision [31][32][33] motivates our approach for ZS-SBIR. However, the decomposition of samples in content and style in our case, is with the sole purpose of better retrieval for ZS-SBIR and thus differs significantly from other work in literature, which we will discuss later.

## III. Proposed StyleGuide for ZS-SBIR

We explain our *StyleGuide* framework in details in this section. We propose to generate fake image features by fusing the sketch-query content and individual styles of the database images and utilize these fake images to retrieve relevant images from the database. With such style-guided fake image generation, *StyleGuide* can reduce the domain difference and image-specific variations, which results in improved retrieval performance. Towards that goal, we design two modules, (1) content-style decomposition module; and (2) content-style fusion module to generate fake image features. We will now describe these modules in details in the following sub-sections. Figure 1 illustrates the proposed approach.

**Notations:** The available sketch and image data for training are $\mathcal{S}_{train} = \{\mathbf{S}_i, l_i^{(S)}\}_{i=1}^{N_S}$ 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 $l_i^{(S)}$, $l_i^{(I)} \in \mathcal{Y}_{seen}$, $\forall i$, where $\mathcal{Y}_{seen}$ represents the set of training classes (seen classes). In contrast to the existing literature, no pairing [6] or same index [7] assumptions for training data are considered in our work. The testing sketch and image data for ZS-SBIR are, $\mathcal{D}_{sketch} = \{\mathbf{S}_j, l_j^{(S)}\}_{j=1}^{M_S}$ and $\mathcal{D}_{image} = \{\mathbf{I}_j, l_j^{(I)}\}_{j=1}^{M_I}$ respectively, where $l_j^{(S)}$, $l_j^{(I)} \in \mathcal{Y}_{unseen}$ and $\mathcal{Y}_{unseen} \cap \mathcal{Y}_{seen} = \phi$. For generalized ZS-SBIR, $\mathcal{D}_{image} = \{\mathbf{I}_j, l_j^{(I)}\}_{j=1}^{M_I^g}$, where $l_j^{(I)} \in \mathcal{Y}_{seen} \cup \mathcal{Y}_{unseen}$. Obviously, the labels of the images and sketches are not available to the algorithm during testing.

**Feature representation:** We obtain data features using pre-trained convolutional networks (CNN), by separately fine-tuning them with $\mathcal{S}_{train}$ and $\mathcal{I}_{train}$. These fine-tuned networks are kept fixed for the rest of the training.

### A. Content-Style decomposition module

The CNN-training is based on label-based loss for classification [34]; however, the extracted features contain domain-specific information of data, since the fine-tuning is performed separately on sketches and images. We aim to decompose these features into domain-independent *content* representation and *style*, which contains the domain-dependent part and residual data-specific information, using the proposed content-style decomposition module. Though this decomposition is inspired by [30], there are important differences which are explained later. Here, we learn two encoders $E_S$ and $E_I$ for sketches and images respectively, to encode the content information as

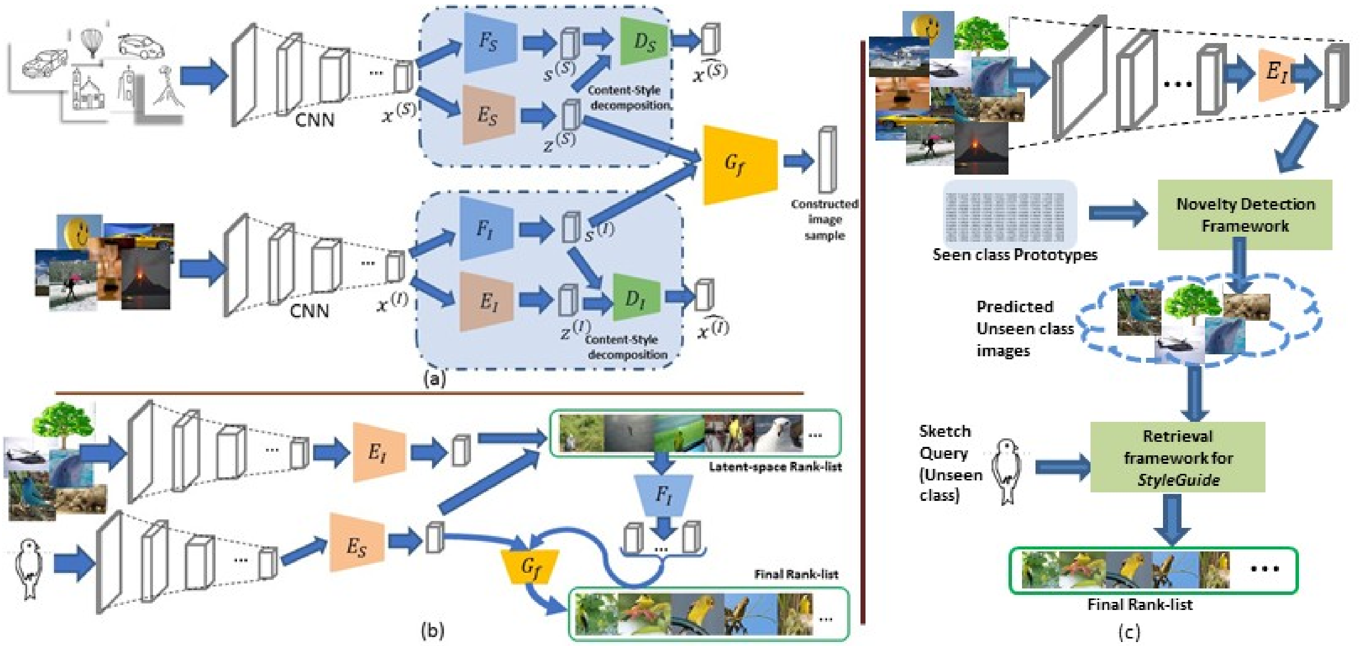

Fig. 1. (a) Illustration of the proposed *StyleGuide* framework for (a) Training and (b) Retrieval. (c) depicts the retrieval workflow using proposed novelty detetction module CDP-ND$_\text{h}$ in addition with *StyleGuide* for generalized ZS-SBIR.

$\mathbf{z}_i^{(m)} = E_m(\mathbf{x}_i^{(m)})$, where $m \in \{S, I\}$ in a shared-space. As in ZSL [35], to ensure better generalization of the model towards the unseen categories, we enforce these shared 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, $d(\mathbf{z}^{(m)}, h(l^{(m)}))$ represents the distance between the extracted content and the category-embedding. Thus, the latent-space representations are learned with distance-based cross-entropy loss on $E_S$ and $E_I$ as,

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

Such category-embedding guided learning of latent-space embeds unseen class samples close to semantically similar seen classes, which minimizes the cross-domain distance between unseen class samples of the same category.

In the second stage of the decomposition, styles present in $\mathbf{x}^{(m)}, m \in \{S, I\}$ is captured. Following [30], 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 $D_m(\mathbf{z}^{(m)}, \mathbf{s}^{(m)}), m \in \{S, I\}$ to reconstruct back the corresponding features, $\mathbf{x}^{(m)}$. Thus, the joint loss function to be minimized at this stage is

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

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

### B. Content-Style fusion module

To generate style-guided image using the query content and search-set image styles, we design a content-style fusion module. A concatenation-based fusion network $G_f$ is designed to combine cross-domain content-style features for meaningful fake image feature generation. We fix the decomposition module of *StyleGuide* and 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 $l_i^{(S)} \neq l_i^{(I^-)}$ to train $G_f$. Generated image features $\hat{\mathbf{x}}^{(I)} = G_f(\mathbf{z}^{(S)}, \mathbf{s}^{(I)})$ follow a margin-based categorical similarity by minimizing the ranking loss function,

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

where, margin $M$ is set experimentally. Essentially, $G_f$ generates fake image features closer to other *real* features from the same class (image matching problem). Proposed fusion module contains a concatenation layer followed by two fc layers with ReLU activation.

### C. Retrieval methodology

In this work, we address category-based ZS-SBIR problem [5][6], where the evaluation involves matching only the category of query sketch and retrieved images.

**a) Retrieval at the Latent Space:** Given a sketch query $\mathbf{x}_q^{(S)}$ and a database of images $\mathcal{D}_{image} = \{\mathbf{I}_{te}\}$, we construct initial rank-list based on the ascending Euclidean distance between their content-information ($\mathbf{z}_q^{(S)}$ and $\mathbf{z}_{te}^{(I)}$) at the latent-space.

**b) Retrieval at the Image Space:** In the second step of retrieval, we select top-K images from the intial rank-list to construct the pruned rank-list $\mathcal{R}_{latent}$. We further fuse $\mathbf{z}_q^{(S)}$ and $\{\mathbf{s}_{te}^{(I)}\}_{te=1}^{K}$ using $G_f$ to generate K-fake image features as $\{\hat{\mathbf{x}}_{te}^{(I)}\}_{te=1}^{K}$. Such fusion eliminates the domain and instance-specific variations and thus the Euclidean distance between $\hat{\mathbf{x}}_{te}^{(I)}$ and $\mathbf{x}_{te}^{(I)}$ reduces further in case they belong to the same category. On the basis of this newly computed distance $d(\hat{\mathbf{x}}_{te}^{(I)}, \mathbf{x}_{te}^{(I)})$, we re-rank $\mathcal{R}_{latent}$ to obtain an improved rank-list $\mathcal{R}_{final}$. Even though such a retrieval mechanism appers to be similar as content-matching, we observe significant performance difference by using such image-space retrieval (details in Experiments section), which is consistent with results (better image-domain matching compared to semantic-space matching) reported in [8] for ZSL.

### D. Difference with Existing Work

**Difference with [6]:** CVAE and CAAE are two ZS-SBIR methods, which generate image features from a given sketch for retrieval. However, (1) our model works for unpaired image-sketch data, overcoming CVAE's requirement of paired data; and (2) *StyleGuide* fuses image-specific styles with sketch-content and uses these fake images for retrieval. In contrast, CVAE generates samples from Gaussian noise and employs a voting methodology for retrieval.

**Difference with [30]:** [30] proposes a general-purpose content-style disentanglement technique, whereas we design the decomposition for the final goal of retrieval. *StyleGuide* uses a distance-based cross-entropy loss in contrast to the class-probability based classification loss as in [30]. Our *style*-definition is also significantly different. We encode the domain-specific knowledge, as well as intra-class variations in the style-vectors, whereas [30] encodes the class-independent information as style.

## IV. Proposed Approach for Generalized ZS-SBIR

In real scenarios, the database may contain images from both seen and unseen classes, which makes the problem much more challenging. The experimental protocol for generalized ZS-SBIR [7] is based on this intuition, where the sketch from an unseen class is queried against a search set of images from both seen and unseen classes. As expected, we observe in [7] and also from our experiments, that the retrieval accuracy for all algorithms drop significantly in this case. Clearly, the presence of images from seen classes in the database increases the chances of confusion which results in degraded performance (more analysis in the Experiments section).

With the prior knowledge that the query sketch belongs to the set of unseen classes, it is intuitive that, an effective method for detecting which search-set image sample belongs to the seen and unseen (novel) classes can be useful in such retrieval scenario. Identification of images belonging to seen or novel classes have also been used successfully to improve the performance of GZSL for classification [26][27]. With this motivation, we propose a simple, yet effective approach, namely CDP-ND (cross-domain prototype based novelty detection), which successfully utilizes both the seen image and sketch data to infer whether an image in the database ($\mathcal{D}_{image}$) belongs to the seen or novel classes. Thus the query sketch needs to be searched only against the newly constructed set $\mathcal{D}_{image}^{(u)}$, consisting of images from the unseen classes. We propose two variants of the CDP-ND - (1) CDP-ND$_\mathrm{s}$: In this soft version of CDP-ND, we use the computed novelty measure to appropriately weigh the similarity scores of the query to the database images; (2) CDP-ND$_\mathrm{h}$: In this hard version of CDP-ND, we use the threshold-based computation of novelty measure to determine whether the database image should be compared with the query or not. Now, we discuss the different steps of the proposed CDP-ND in details.

**Cross-domain Prototype Computation:** Towards the goal of detecting unseen class images in the search set, we compute a set of class-prototypes as, $\mathcal{P}_{seen} = \{\mathbf{p}_j, j \in \mathcal{Y}_{seen}\}$, which are stored in the memory after training. In this work, we exploit the sketch-image cross-domain similarity information to compute these seen class prototypes. The class prototypes are expected to be a reasonable representation of all the sketch-image samples in the latent space. Thus, we propose to compute the prototypes as the mean content vector of the *good* training samples (both image and sketch). The measure of such *goodness* condition of a sample is evaluated in terms of the retrieval accuracy of the corresponding sample in the latent space. Thus, the set of good samples is constructed as

$$\mathbb{C}_{good} = \{\mathbf{z}_i^{(m)} | AP(\mathbf{z}_i^{(m)}, \mathcal{R}_{latent}, M) \geq \delta\} \qquad (4)$$

Here, $AP(\mathbf{z}_i^{(m)}, \mathcal{R}_{latent}, M)$ represents the average precision measured on the top-$M$ retrieved images in $\mathcal{R}_{latent}$ on the basis of sorted Euclidean distance with $\mathbf{z}_i^{(m)}$. $\delta$ acts as a threshold for selecting a sample on the basis of its *good*ness. Thus the set of *good* samples of class-$j$ is given as

$$\mathcal{C}_j = \{\mathbf{z}_i^{(m)} | l_i^{(m)} = j, \text{ and } \mathbf{z}_i^{(m)} \in \mathbb{C}_{good}\}, m \in \{S, I\} \qquad (5)$$

from which the prototype of class-$j$ is computed as

$$\mathbf{p}_j = mean(\mathcal{C}_j) \qquad (6)$$

Here, the *good* training examples are only used to compute $\mathbf{p}_j$ instead of all samples of $j^{th}$ class, so that the outliers do not adversely effect the computed prototypes. An advantage of the proposed novelty detection approach is that the training process need not be modified. The distance of each database image with these prototypes are used to weigh the similarity score in CDP-ND$_\mathrm{s}$ or to classify them as belonging to seen or novel classes in CDP-ND$_\mathrm{h}$ as described below.

**GZS-SBIR using proposed CDP-ND$_\mathrm{s}$:** In this variant, the query sketch is compared with all the database images, but each image is given a weight corresponding to its chance of belonging to a seen or novel class. The intuition is that if a database image belonging to an unseen class is incorrectly labeled as belonging to a seen class using the prototypes, that image is not removed from the comparison set and can still be retrieved if the similarity score is considerably high.

Towards that goal, we compute similarity scores between the database image contents $\mathbf{z}_{te}^{(I)}$, s.t., $\mathbf{I}_{te} \in \mathcal{D}_{image}$ and the saved class-prototypes $\mathcal{P}_{seen}$ as

$$score(\mathbf{z}_{te}^{(I)}, \mathbf{p}_j) = cosine\_similarity(\mathbf{z}_{te}^{(I)}, \mathbf{p}_j) \quad (7)$$

If $\mathbf{I}_{te}$ belongs to one of the seen classes, then it should have a high similarity score with that particular class prototype. On the other hand, if it belongs to a novel class, it will have low similarity values with all the seen class prototypes. Hence, we consider the maximum similarity score as the prototypical similarity of the test sample, which is evaluated as

$$score_p(\mathbf{z}_{te}^{(I)}) = max_j \; score(\mathbf{z}_{te}^{(I)}, \mathbf{p}_j) \quad (8)$$

Let the similarity score between this image sample and the test sketch query $\mathbf{S}_q$ be given by

$$score_q(\mathbf{z}_{te}^{(I)}, \mathbf{z}_q^{(S)}) = cosine\_similarity(\mathbf{z}_{te}^{(I)}, \mathbf{z}_q^{(S)}) \quad (9)$$

The final novelty-weighted similarity score for $\mathbf{z}_{te}^{(I)}$ used for retrieval is computed as (arguments dropped for clarity)

$$score_{final} = score_q(1 - score_p) \quad (10)$$

We observe that the prototypical similarity score is used to softly weigh the content-based similarity score. If a database image is likely to belong to one of the seen classes, $score_p$ is higher, which in turn will give a lower weight to $score_{final}$. On the other hand, if the database image is likely to belong to a novel class, $score_p$ is significantly lower, which results in a comparatively higher weight for $score_{final}$. Hence, CDP-ND$_\text{s}$ enables the algorithm to retrieve the unseen class images with a high probability compared to the seen class image samples present in the database.

**GZS-SBIR using proposed CDP-ND$_\text{h}$:** In this variant, we propose to use the similarity of each database image with the seen class prototypes $\mathcal{P}_{seen}$ to completely separate out the data from the seen and novel classes. Thus the query sketch (which belong to an unseen class) can be compared to a significantly smaller set of images which are likely to belong to the unseen classes.

Before retrieval, we compute the cosine-similarity between the content-encoding $\mathbf{z}_{te}^{(I)}$ of database image samples $\mathbf{I}_{te} \in \mathcal{D}_{image}$ and saved class-prototypes $\mathcal{P}_{seen}$ as in (7). Since the unseen class images are expected to have lower similarity compared to the seen class samples with the prototypes, based on the similarity scores, the refined search-set $\mathcal{D}_{image}^{(u)} \subset \mathcal{D}_{image}$ is constructed as,

$$\mathcal{D}_{image}^{(u)} = \{\mathbf{I}_{te} | score(\mathbf{z}_{te}^{(I)}, \mathbf{p}_j) < \epsilon, \forall j \in \mathcal{Y}_{seen}\} \quad (11)$$

Here, $\epsilon$ is an experimental hyper-parameter which acts as the threshold to detect the novel classes. Thus, once $\mathcal{D}_{image}^{(u)}$ is constructed, the same retrieval methodology (as described in Section III-C) is used on $\mathcal{D}_{image}^{(u)}$ for final retrieval.

## V. 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 datasets used in this work.

### A. Datasets and Implementation Details

**The Sketchy Dataset** [9] is a collection of 75,471 sketches and 12,500 images from 125 classes. We used the extended dataset [3] containing additional 60,502 images. For ZS-SBIR, experiments are performed on two different data-splits 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. In contrast, [6] proposes another data-split (**Split 2**), where 21 categories which are not part of the ImageNet [36] are selected as unseen and the rest are used for training. This split ensures that the unseen classes are truly unknown to the model, even though pre-trained models are used for feature representation.

**TU-Berlin dataset** [10] contains 20,000-sketches from 250 categories and is extended by [3] 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 [5][7] for ZS-SBIR.

**Implementation Details:** *StyleGuide* is implemented in TensorFlow [37]. 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 training images and sketches and fc7-features are considered as $\mathbf{x}^{(m)}, m \in \{S, I\}$. The category-name embeddings $h(.)$ (200-d) are extracted using pre-trained GloVe [38] model. $\mathbf{z}^{(m)}$ and $\mathbf{s}^{(m)}, m \in \{S, I\}$ are restricted to be of 200-d and 100-d, respectively. Adam optimizer with $\beta_1 = 0.5, \beta_2 = 0.999$, a learning rate of $10^{-3}$ for content-style decomposition and $10^{-4}$ for fusion with a batch-size of 64 and 32 for Sketchy and TU-Berlin, respectively are used.

### B. Experients for ZS-SBIR protocol

We compare *StyleGuide* with state-of-the-art SBIR methods and ZSL algorithms. For direct comparison with results in literature, we perform experiments on both splits of Sketchy and standard split of TU-Berlin.

Table I reports the results on Sketchy (Split 1) and TU-Berlin (standard ZS-SBIR split as in [5][7]) in terms of MAP@all and Precision@100. All the results for the other approaches are taken from [7]. We observe that the proposed *StyleGuide* outperforms the state-of-the-art on Sketchy. For TU-Berlin, we achieve second best performance, which is only less than [7] and better than all the others. We perform additional experiments on Sketchy Split 2. Table II reports MAP@200 and Precision@200 for the same. Here, the results of the other approaches are directly taken from [6]. Our search-set specific style-guided retrieval clearly outperforms all approaches, including both noise-based generation methods CVAE and CAAE, by significant margins.

TABLE I

Comparison (Precision@100, MAP@all) of StyleGuide with ZS-SBIR methods on Sketchy (Split 1) and TU-Berlin.

| Type | Methods | Sketchy (Split 1) | | | TU-Berlin | | |
|---|---|---|---|---|---|---|---|
| | | Precision@100 | MAP@all | Retrieval time (sec.) | Precision@100 | MAP@all | Retrieval time (sec.) |
| SBIR | Softmax Baseline | 0.172 | 0.114 | $3.5 \times 10^{-1}$ | 0.143 | 0.089 | $4.3 \times 10^{-1}$ |
| | Siamese CNN [14] | 0.175 | 0.132 | $5.7 \times 10^{-3}$ | 0.141 | 0.109 | $5.9 \times 10^{-3}$ |
| | SaN [39] | 0.125 | 0.115 | $4.8 \times 10^{-2}$ | 0.108 | 0.089 | $5.5 \times 10^{-2}$ |
| | GN Triplet [9] | 0.296 | 0.204 | $9.1 \times 10^{-2}$ | 0.253 | 0.175 | $1.9 \times 10^{-1}$ |
| | 3D shape [40] | 0.078 | 0.067 | $7.8 \times 10^{-3}$ | 0.067 | 0.054 | $7.2 \times 10^{-3}$ |
| | DSH [3] | 0.231 | 0.171 | $6.1 \times 10^{-5}$ | 0.189 | 0.129 | $7.2 \times 10^{-5}$ |
| | GDH [17] | 0.259 | 0.187 | $7.8 \times 10^{-5}$ | 0.212 | 0.135 | $9.6 \times 10^{-5}$ |
| ZSL | CMT [41] | 0.102 | 0.087 | $2.8 \times 10^{-2}$ | 0.078 | 0.062 | $3.3 \times 10^{-2}$ |
| | DeViSE [42] | 0.077 | 0.067 | $3.6 \times 10^{-2}$ | 0.071 | 0.059 | $3.2 \times 10^{-2}$ |
| | SSE [43] | 0.161 | 0.116 | $1.3 \times 10^{-2}$ | 0.121 | 0.089 | $1.7 \times 10^{-2}$ |
| | JLSE [44] | 0.185 | 0.131 | $1.5 \times 10^{-2}$ | 0.155 | 0.109 | $1.4 \times 10^{-2}$ |
| | SAE [20] | 0.293 | 0.216 | $2.9 \times 10^{-2}$ | 0.221 | 0.167 | $3.2 \times 10^{-2}$ |
| | FRWGAN [45] | 0.169 | 0.127 | $3.2 \times 10^{-2}$ | 0.157 | 0.110 | $3.9 \times 10^{-2}$ |
| | ZSH [46] | 0.214 | 0.159 | $5.9 \times 10^{-5}$ | 0.177 | 0.141 | $7.6 \times 10^{-5}$ |
| ZS-SBIR | ZSIH [5] | 0.342 | 0.258 | $6.7 \times 10^{-5}$ | 0.294 | 0.223 | $7.7 \times 10^{-5}$ |
| | ZS-SBIR [6] | 0.284 | 0.196 | $9.6 \times 10^{-2}$ | 0.001 | 0.005 | $1.2 \times 10^{-1}$ |
| | SEM-PCYC [7] | 0.463 | 0.349 | $1.7 \times 10^{-3}$ | 0.426 | 0.297 | $1.9 \times 10^{-3}$ |
| | **StyleGuide** | **0.4842** | **0.3756** | $1.6 \times 10^{-2}$ | **0.3551** | **0.2543** | $5.7 \times 10^{-2}$ |

TABLE III

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 for generalized ZS-SBIR protocol.

| Methods | Sketchy (Split 1) | | | TU-Berlin | | |
|---|---|---|---|---|---|---|
| | Precision@100 | MAP@all | Retrieval time (sec.) | Precision@100 | MAP@all | Retrieval time (sec.) |
| ZSIH [5] | 0.296 | 0.219 | $6.7 \times 10^{-5}$ | 0.218 | 0.142 | $7.7 \times 10^{-5}$ |
| SEM-PCYC [7] | 0.364 | 0.307 | $1.7 \times 10^{-3}$ | 0.298 | 0.192 | $2.0 \times 10^{-3}$ |
| **StyleGuide** | **0.3811** | **0.3307** | $1.6 \times 10^{-2}$ | **0.2264** | **0.1488** | $5.7 \times 10^{-2}$ |
| StyleGuide + Openmax-based ND [24] | 0.3771 | 0.3382 | $7.1 \times 10^{-2}$ | 0.2295 | 0.1501 | $1.8 \times 10^{-1}$ |
| StyleGuide + AE-based ND [28] | 0.4030 | 0.3500 | $3.9 \times 10^{-2}$ | 0.2346 | 0.1533 | $3.2 \times 10^{-2}$ |
| **StyleGuide + CDP-ND$_{\text{s}}$** | 0.4317 | 0.3529 | $8.4 \times 10^{-2}$ | 0.2764 | 0.1966 | $9.3 \times 10^{-2}$ |
| **StyleGuide + CDP-ND$_{\text{h}}$** | **0.4533** | **0.3671** | $5.0 \times 10^{-2}$ | **0.2911** | **0.2146** | $6.7 \times 10^{-2}$ |

TABLE II

Comparison (Precision@200 and MAP@200) with existing ZS-SBIR methods on Sketchy (Split 2) data.

| Type | Evaluation methods | Precision | MAP |
|---|---|---|---|
| | | @200 | @200 |
| SBIR methods | Baseline | 0.106 | 0.054 |
| | Siamese-1 [47] | 0.243 | 0.134 |
| | Siamese-2 [15] | 0.251 | 0.149 |
| | Coarse-grained Triplet [9] | 0.169 | 0.083 |
| | Fine-grained Triplet | 0.155 | 0.081 |
| | DSH [3] | 0.153 | 0.059 |
| ZSL methods | Direct Regression | 0.066 | 0.022 |
| | ESZSL [21] | 0.187 | 0.117 |
| | SAE [20] | 0.238 | 0.136 |
| ZS-SBIR | CAAE [6] | 0.260 | 0.156 |
| | CVAE [6] | 0.333 | 0.225 |
| | **StyleGuide** | **0.4001** | **0.3581** |

### *C. Experiments for Generalized ZS-SBIR protocol*

We report the retrieval results of *StyleGuide* (without the proposed novelty detection) for generalized ZS-SBIR in the top half of Table III. The results for the other algorithms are directly taken from [7]. We observe *StyleGuide* achieves state-of-the-art results for Sketchy dataset, and performs slightly less than [7] for TU-Berlin.

We also observe that retrieval MAP for all the approaches is significantly lower compared to ZS-SBIR (Table I). The presence of seen class images in the search set makes generalized ZS-SBIR even more challenging. We notice the same in Fig. 2 displaying top-5 retrieved images for sample sketches for both ZS-SBIR (left) and generalized ZS-SBIR (right). The wrongly retrieved images for generalized ZS-SBIR are mostly from the set of seen classes. Motivated by this, we perform the unseen class image detection prior to retrieval, so that the seen class images in the database can either be weighted less or removed from comparison.

**Generalized ZS-SBIR results with novelty detection:** Here, we explore whether novelty detection techniques can be utilized to improve the retrieval performance in generalized ZS-SBIR by trying to answer the following questions.

1) Since our goal is to identify if a given image belongs to the unseen class or not, can standard novelty detection techniques developed solely for images be used to improve the retrieval results?

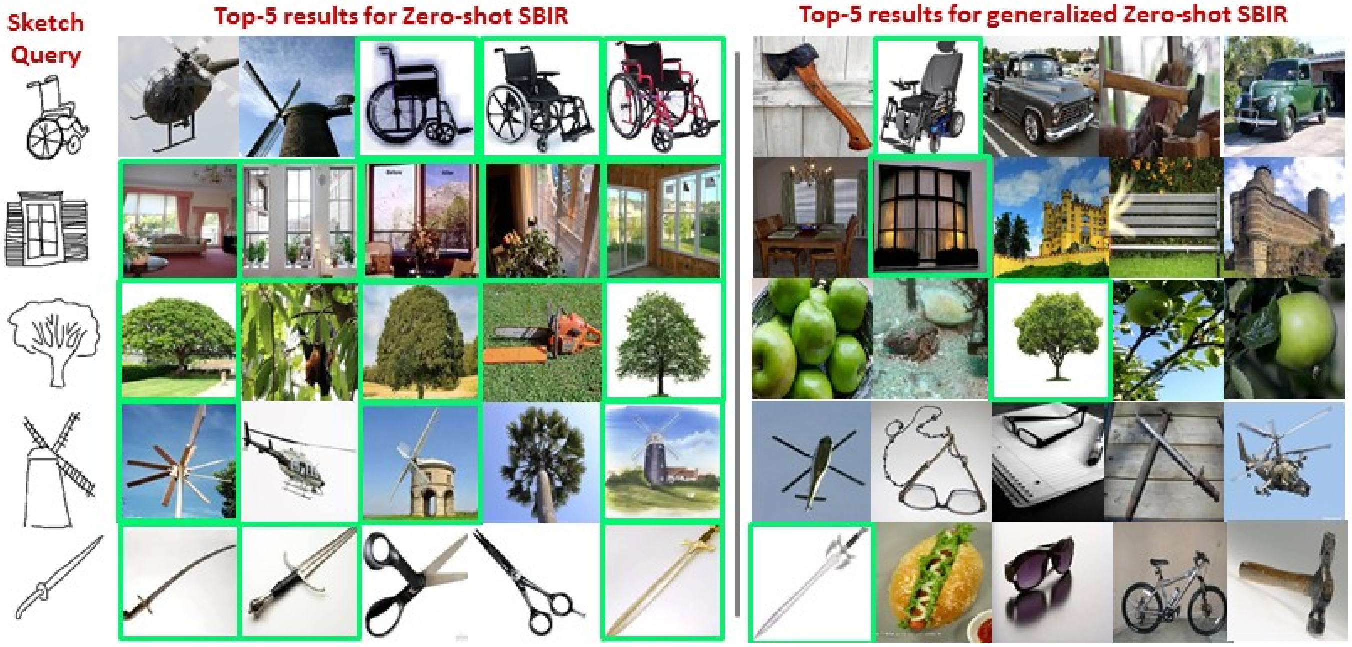
Sketch Query
Top-5 results for Zero-shot SBIR
Top-5 results for generalized Zero-shot SBIR

Fig. 2. Top-5 retrieved images for same sketch query for ZS-SBIR (left) and generalized ZS-SBIR (right). The correctly retrieved images are highlighted with a green border in both cases.

2) How does CDP-ND which utilises both the sketch and image domains compare with these techniques?

To answer the first question, we choose the following two algorithms which have been developed for multi-class novelty detection for image data, namely (1) Openmax [24], and (2) AE-based novelty detection [28].

(1) *Openmax* [24]: In this method, for a test image sample, the softmax probabilities of the known classes are re-weighted according to the statistics of the training samples near the boundary regions of the respective class. These statistics are evaluated by fitting Weibull distribution on such samples. Such re-weighted probability scores open-up the detection probability of *unknown unknown* samples in the test set. We follow the same re-weighting strategy using the seen class probabilities $p(l^{(I)}|\mathbf{x}_{te}^{(I)})$ (as in (1)). During retrieval, only the predicted unseen class images are considered as $\mathcal{D}_{image}^{(u)}$ and used as the effective search set as in Section III-C.

(2) *Auto-encoder-based novelty detection* [28]: In this algorithm, an auto-encoder (AE) architecture has been modified to obtain a novelty score associated with an image, based on the intuition that the reconstructed image should be significantly degraded in quality in case the input image sample belongs to a class previously unseen to the AE. We use model-2 (M-2) mentioned in this work which uses the degraded reconstruction criteria when the input image is paired with incorrect class-attribute and gives better results as compared to the alternate model (M-1). Using this approach, we construct $\mathcal{D}_{image}^{(u)}$ and then use the retrieval methodology of *StyleGuide*.

The retrieval performance of *StyleGuide* combined with these two novelty detection methods for generalized ZS-SBIR is reported in the bottom half of Table III. We observe noticable improvement in MAP for both the datasets using [28] as the novelty-detector.

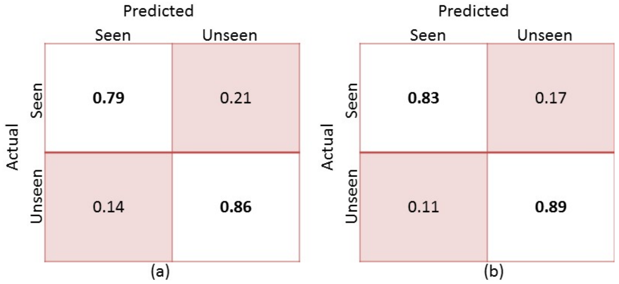

Fig. 3. Novelty detection accuracy using (a) AE-based novelty detector [28], (b) proposed CDP-ND$_{\mathrm{h}}$ on the search-set images $\mathcal{D}_{image}$ of Sketchy Split 2.

**Generalized ZS-SBIR results with CDP-ND:** Now to answer the second question, first we evaluate the accuracy of the proposed CDP-ND to classify a given image as seen or novel as compared to the AE-based novelty-detector [28]. We observe from the confusion matrix in Fig. 3 that CDP-ND$_{\mathrm{h}}$ performs better than the AE based approach on the search-set images of Sketchy Split 2. The pattern is similar for the other variant CDP-ND$_{\mathrm{s}}$. This motivates us to use the proposed approaches for novelty detection for generalized ZS-SBIR.

Finally, we evaluate the effectiveness of the proposed CDP-ND technique for generalized ZS-SBIR and report the results of both the variants in the last part of Table III. We observe that both the proposed variants help to considerably boost the retrieval performance, and interestingly, the hard thresholding variant CDP-ND$_{\mathrm{h}}$ performs better than the softer weighting version CDP-ND$_{\mathrm{s}}$. We also observe that *StyleGuide* with CDP-ND$_{\mathrm{h}}$ outperforms the state-of-the-art for both the datasets.

TABLE IV
ANALYSIS WITH DIFFERENT BASELINES.

| Description (ZS-SBIR) | 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 |
| **StyleGuide** | **0.3581** |
| Description (generalized ZS-SBIR) | MAP@200 |
| StyleGuide | 0.2872 |
| StyleGuide+CDP-ND$_\text{h}^{le}$ | 0.2935 |
| StyleGuide+CDP-ND$_\text{h}^{ptn}$ | 0.3139 |
| StyleGuide+CDP-ND$_\text{h}^{image}$ | 0.3102 |
| **StyleGuide+CDP-ND$_\text{h}$** | **0.3317** |

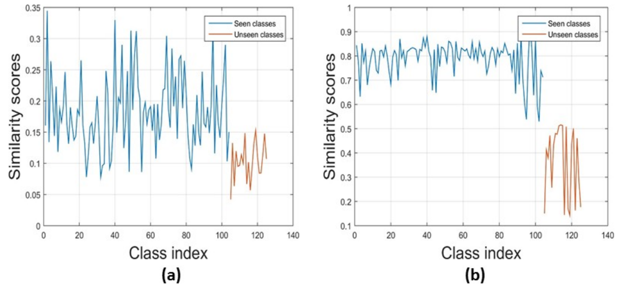

Fig. 4. Comparison of $score_p$ for seen and unseen classes using prototype-variants: (a) CDP-ND$_\text{h}^{le}$, and (b) CDP-ND$_\text{h}$ on Sketchy Split 2. The scores for unseen classes are highlighted with *red*-color.

## VI. Analysis and Discussion

We perform detail analysis of our framework in this section. Specifically, we discuss different variants of the framework to better understand the usefulness of each network components. All the analysis are performed on Sketchy extended (Split-2) dataset, unless specified otherwise.

**1) Variations for ZS-SBIR:** We analyze *StyleGuide* by developing different baselines by modifying individual modules or the loss-terms. The results obtained are reported in Table IV. In B1, the images are retrieved only through the content-matching of sketches ($\mathbf{z}^{(S)}$) and images ($\mathbf{z}^{(I)}$) in $\mathcal{R}_{latent}$. B2 reports the results when $E_I$ and $E_S$ are trained using standard cross-entropy loss using class labels instead of category-word vectors. The retrieval accuracy is lower in such case, since such loss function does not contain any semantic information about the categories. In B3, final retrieval accuracy is measured on the basis of fused similarity-values in both *content* and *image*-space. However, to our surprise, this did not yield any improvement over proposed *StyleGuide*-retrieval. Possible reason may be the style-guided reconsruction of images may not contain any complimentary information to boost the performance further. Finally, in B4, we generate single fake image from the query by fusing it with randomly selected style from the database, instead of an image-specific style in $\mathcal{R}_{latent}$. We perform retrieval on the basis of similarity of this single generated image with all the images in search set and observe a considerable drop in the performance which justifies the importance of selection of styles. The full proposed model, with style-based final ranking, produces the best result.

**2) Variations for generalized ZS-SBIR:** In the proposed novelty detection framework, we compute the prototypes for each seen class by utilizing good samples from both the sketch and image domains. We further evaluate the retrieval performance using different variants for computing the seen class-prototypes: (1) Label embeddings (LE): Since the latent-space construction in our approach (equation (2)) is based on the label embeddings of the seen classes, we can take these embeddings $(h(l^I))$ of seen class labels to be the prototypes, i.e. $\mathbf{p}_j = h(l_j^{(I)})$. This variant is denoted as StyleGuide+CDP-ND$_\text{h}^{le}$. (2) Prototypical Networks (PTN): [48] describes an end-to-end deep network for few-shot learning applications based on episodic training to obtain prototypes of the training classes. In this variant, we use PTN to generate the seen class image prototypes. However, instead of selecting a set of training classes in each episode, we follow the standard mini-batch training and the loss on the image samples in the validation set is minimized. After training, the class prototype is evaluated as, $\mathbf{p}_j = \sum_{\mathcal{P}} \mathbf{z}^{(I)}$, s.t., $\mathcal{P} = \{\mathbf{z}^{(I)} | l^{(I)} = j\}$. This variant is denoted as StyleGuide+CDP-ND$_\text{h}^{ptn}$ (3) Single-domain based prototypes: To evaluate whether using both domain data helps in computing better prototypes, in this variant, we use only the good image data to construct the $\mathbf{C}_{good}$ in (4). The class prototypes are evaluated accordingly. This variant is denoted as StyleGuide+CDP-ND$_\text{h}^{image}$.

The evaluation results using all these variants are reported in Table IV. We observe that the proposed approach of computing the seen class prototypes utilizing both sketch and image data outperforms all the other variants. Fig. 4 shows the $score_p$-value (8) averaged over all the samples in a particular class present in $\mathcal{D}_{image}$ for generalized ZS-SBIR for both CDP-ND$_\text{h}^{le}$ and CDP-ND$_\text{h}$. We observe a clear discrimination in $score_p$ for seen and unseen classes for CDP-ND$_\text{h}$ as compared to CDP-ND$_\text{h}^{le}$, which justifies its better detection and retrieval accuracy. The other variations of CDP-ND$_\text{h}$ display a similar pattern as CDP-ND$_\text{h}^{le}$ and hence are not included here.

## VII. Conclusion

In this work, we proposed a novel approach based on content-style decomposition, termed as *StyleGuide* for the challenging task of ZS-SBIR. We also extended our framework for generalized ZS-SBIR protocol using a novel unseen-class image detection mechanism CDP-ND$_\text{h}$, which provides a significant boost to the performance of *StyleGuide* in such generalized setting. Using such a detection method, our approach can seamlessly be applied to cases where even the a-priori knowledge of the unseen class query sample is not available. Extensive experiments and analysis has been performed and proposed *StyleGuide* in combination with CPD-ND$_\text{h}$ is observed to outperform state-of-the-art retrieval methods on two large-scale datasets.

## References

[1] 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.

[2] 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.

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

[4] 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*, vol. 27, no. 9, pp. 4410–4421, 2018.

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

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

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

[8] 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.

[9] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," *ACM TOG*, vol. 35, no. 4, pp. 1–12, 2016.

[10] E. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM TOG*, vol. 31, no. 4, pp. 1–10, 2012.

[11] T. Dutta and S. Biswas, "Style-guided zero-shot sketch-based image retrieval," in *BMVC*, 2019.

[12] R. Hu and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *CVIU*, vol. 117, no. 7, pp. 790–806, 2013.

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

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

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

[16] 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*, vol. 71, pp. 77–87, 2018.

[17] 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.

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

[19] 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.

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

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

[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] 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.

[24] A. Bendale and T. E. Bolt, "Towards open set deep networks," in *CVPR*, 2016.

[25] Q. Yu and K. Aizawa, "Unsupervised out-of-distribution detection by maximum classifier discrepancy," in *ICCV*, 2019.

[26] D. Mandal, S. Narayan, S. K. Dwivedi, V. Gupta, S. Ahmed, F. S. Khan, and L. Shao, "Out-of-distribution detection for generalized zero-shot action recognition," in *CVPR*, 2019.

[27] O. Gune, A. More, B. Banerjee, and S. Chaudhuri, "Generalized zero-shot learning using open-set recognition," in *BMVC*, 2019.

[28] S. Bhattacharjee, D. Mandal, and S. Biswas, "Autoencoder based novelty detection for generalized zero-shot learning," in *ICIP*, 2019.

[29] 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.

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

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

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

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

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

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

[36] 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*, vol. 115, no. 3, pp. 211–252, 2015.

[37] 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.

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

[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*, vol. 122, no. 3, pp. 411–425, 2017.

[40] 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*, vol. 8, no. 10, pp. 998–1009, 2015.

[41] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Neur-IPS*, 2013.

[42] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: a deep visual-semantic embedding model," in *Neur-IPS*, 2013.

[43] 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*, vol. 24, no. 12, pp. 4766–4779, 2015.

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

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

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

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

[48] J. Snell, K. Swerky, and R. Zemel, "Prototypical networks for few-shot learning," in *Neur-IPs*, 2017.