

Doodle to Search: Practical Zero-Shot Sketch-based Image Retrieval

Sounak Dey*, Pau Riba*, Anjan Dutta, Josep Lladós
 Computer Vison Center, UAB, Spain
 {sdey, priba, adutta, josep}@cvc.uab.cat

Yi-Zhe Song
 SketchX, CVSSP, University of Surrey, UK
 y.song@surrey.ac.uk

Abstract

In this paper, we investigate the problem of zero-shot sketch-based image retrieval (ZS-SBIR), where human sketches are used as queries to conduct retrieval of photos from unseen categories. We importantly advance prior arts by proposing a novel ZS-SBIR scenario that represents a firm step forward in its practical application. The new setting uniquely recognizes two important yet often neglected challenges of practical ZS-SBIR, (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval. We first contribute to the community a novel ZS-SBIR dataset, QuickDraw-Extended, that consists of 330,000 sketches and 204,000 photos spanning across 110 categories. Highly abstract amateur human sketches are purposefully sourced to maximize the domain gap, instead of ones included in existing datasets that can often be semi-photorealistic. We then formulate a ZS-SBIR framework to jointly model sketches and photos into a common embedding space. A novel strategy to mine the mutual information among domains is specifically engineered to alleviate the domain gap. External semantic knowledge is further embedded to aid semantic transfer. We show that, rather surprisingly, retrieval performance significantly outperforms that of state-of-the-art on existing datasets that can already be achieved using a reduced version of our model. We further demonstrate the superior performance of our full model by comparing with a number of alternatives on the newly proposed dataset. The new dataset, plus all training and testing code of our model, will be publicly released to facilitate future research[†].

1. Introduction

In the context of retrieval, sketch modality has shown great promise thanks to the pervasive nature of touchscreen devices. Consequently, research on sketch-based image retrieval (SBIR) has flourished, with many great examples addressing various aspects of the retrieval process: fine-

*These authors contributed equally to this work.

[†]<http://dag.cvc.uab.es/doodle2search/>

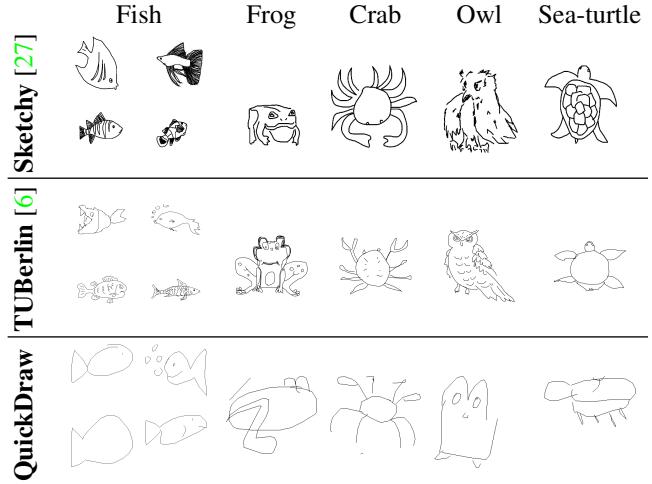


Figure 1. Qualitative comparison of sketch datasets, columns show examples belonging the same class. Sketchy, TUBerlin and QuickDraw datasets orderly contain sketches with increasing level of abstraction. It is worth noting that despite being the most abstract dataset, QuickDraw sketches can still be reliably recognised.

grained matching [37, 30, 24], large-scale hashing [17, 16], cross-modal attention [5, 30] to name a few.

However, a common bottleneck identified by almost all sketch researches is that of data scarcity. Different to photos that can be effortlessly crawled for free, sketches have to be drawn one by one by human being. As a result, existing SBIR datasets suffer in both volume and variety, leaving only less than thousand of sketches per category, with maximum number of classes limited to few hundreds. This largely motivated the problem of zero-shot SBIR (ZS-SBIR), where one wishes to conduct SBIR on object categories without having the training data. ZS-SBIR is increasingly being regarded as an important component in unlocking the practical application of SBIR, since million-scale datasets that have been used to train commercial photo-only systems [4] might not be feasible.

The problem of ZS-SBIR is extremely challenging. It shares all challenges laid out in conventional SBIR: (i) large domain gap between sketch and image, and (ii) high degree

of abstraction found in human sketches as a result of variant drawing skills and visual interpretations. Additionally, it also needs the semantic transference from the seen to unseen categories for the purpose of zero-shot learning. Over and above all, in this paper, we are interested in moving towards the practical adaptation of ZS-SBIR technology. For that, a more appropriate dataset that best capture all these challenges is required.

Therefore, our first contribution is a new dataset to simulate the real application scenario of ZS-SBIR, which should satisfy the following requirements. First, the dataset needs to mimic the real-world abstraction gap between sketch and photo. Such amateur sketches are very different from the ones currently studied by existing datasets, which are either too photo-realistic [7] or produced by recollection of a reference images [27] (Figure 1 offers a comparative example). Second, in order to learn a reliable cross-domain embedding between amateur sketch and photo, the dataset much faithfully capture of a full variety of sketch samples from users having various drawing skills. Our proposed dataset, *QuickDraw-Extended*, contains 330,000 sketches and 204,000 photos in total spanning across 110 categories. In particular, it includes 3,000 amateur sketches per category carefully sourced from the recently released Google Quickdraw dataset [12] – six times more than the next largest. It also has a search space stretching to 166million total comparisons in the test set, compared to *Sketchy-Extended* and *TUBerlin-Extended* with just 10 million and 1.9 million, respectively.

This dataset and the real-world scenario it mimics, essentially make the ZS-SBIR task more difficult. This leads to our second contribution which is a novel cross-domain zero-shot embedding model that addresses all challenges posed by this new setting. Our base network is a visually-attended triplet ranking model that is commonly known in the SBIR community to produce state-of-the-art retrieval performances [37, 30]. To our surprise, just by adopting such a triplet formulation, we can already achieve retrieval performances drastically better than that of the previously reported ZS-SBIR results on commonly used datasets. We attribute this phenomena to previous datasets being too simplistic in terms of the cross-domain abstraction gap and the diversity of sketch samples. This further justifies the necessity of a new practical dataset like ours. We then propose two novel techniques to help learn a better cross-domain transfer model. First, a domain disentanglement strategy is designed to bridge gap between the domains by forcing the network to learn a domain-agnostic embedding, where a *Gradient Reversal Layer* (GRL) [8] encourages the encoder to extract mutual information from sketches and photos. Second, a novel semantic loss to ensure that semantic information is preserved in the obtained embedding. By applying a GRL only to the negative samples at the input of

the semantic decoder helps the encoder network to separate the semantic information of similar classes.

Extensive experiments are first carried out on the two commonly used ZS-SBIR datasets, TU Berlin-Extended [6] and Sketchy-Extended [27]. The results show that the even a reduced version of our model can outperform current state-of-the-arts by a significant margin. The superior performance of the proposed method is further validated on our own dataset, with ablative studies to draw insights towards each of the proposed system components.

2. Related Work

SBIR Datasets. One of the key barriers towards large-scale SBIR research is the lack of appropriate benchmarks. The Sketchy dataset [27] is the mostly used one for this purpose, which contains 75,471 hand-drawn sketches of 12,500 object photos belonging to 125 different categories. Later, Liu *et al.* [17] collected 60,502 natural images from ImageNet [4] in order to fit the task of large-scale SBIR. This dataset having contained highly detailed or less abstract sketches, models trained on Sketchy have high chance of getting collapsed in real life scenario. Two more fine-grained SBIR datasets with paired sketches and images are *shoe* and *chair* datasets which were proposed in [37]. The *shoe* dataset contains altogether 6648 sketches and 2000 photos, whereas, the *chair* dataset altogether contains 297 sketches and photos. However, being fine-grained pairs these two datasets also have similar disadvantages as the Sketchy dataset. TU-Berlin [6] being the other popular dataset originally contains 250 classes of hand-drawn sketches, where each class roughly contains 80 instances. It was extended with real images by [38] for SBIR purposes. This dataset has a lot of confusion regarding the class hierarchy, for an example, swan, seagull, pigeon, parrot, duck, penguin, owl have substantial visual similarity and commonality with standing bird and flying bird which are another separate categories of the TU-Berlin dataset. To obliterate, these difficulties faced by the SBIR works, in this paper, we introduce QuickDraw-Extended dataset, where we take the sketch classes of the Google QuickDraw dataset [12] and provide the corresponding set of images to facilitate the training of large-scale SBIR system.

Sketch-based Image Retrieval (SBIR). The main challenge that most of the SBIR tasks address is bridging the domain gap between sketch and natural image. In literature, these existing methods can be roughly grouped into two categories: *hand crafted* and *cross-modal deep learning* methods. The hand-crafted techniques mostly work with Bag-of-Words representations of sketch and edge map of natural image on top of some off-the-shelf features, such as, SIFT [19], Gradient Field HOG [10], Histogram of Edge Local Orientations [25] or Learned Key Shapes [26]

etc. This domain shift issue is further addressed by cross-domain deep learning-based methods [27, 37], where they have used classical ranking losses, such as, contrastive loss, triplet loss [32] or more elegant HOLEF loss [30] within a siamese like network. Based on the problem at hand, two separated tasks have been identified: (1) *Fine-grained SBIR* (FG-SBIR) aims to capture fine-grained similarities of sketch and photo [15, 27, 37] and (2) *Coarse-grained SBIR* (CG-SBIR) performs a instance level search across multiple object categories [38, 10, 11, 31, 38], which has received a lot of attention due to its importance. Realising the need of large-scale SBIR, some researchers have proposed a variant of cross-modal hashing framework for the same [17, 39], which also showed promising results in SBIR scenario. In contrast, our proposed model overcomes this domain gap by mining the modality agnostic features using a domain loss along with a GRL.

Zero-Shot Sketch-based Image Retrieval (ZS-SBIR). Early works on zero-shot learning (ZSL) were mostly focused on attribute based recognition [14], which is later augmented by another major line that focus on learning a joint embedding space for image feature representation and class semantic descriptor [3, 34, 13, 35, 18]. Depending on the selection of joint embedding space and type of projection function utilised between the visual to semantic space, existing models can be divided into three groups: (i) projected from visual feature space to semantic space [14, 21], (ii) projected from semantic space to the visual feature space [3], and (iii) an intermediate space that both are simultaneously projected to [40]. In contrast to these existing works, our model can be seen as a combination of the first and second groups, where the embedding is on the visual feature space, but asked to additionally recover its embodied semantics with a decoder.

Although SBIR and ZSL have been extensively studied among the research community, very few works have studied their combination. Shen *et al.* [28] propose a multi-modal network to mitigate the sketch-image heterogeneity and enhance semantic relations. Yelamarthi *et al.* [36] resort to a deep conditional generative model, where a sketch is taken as input and learned to generate its photo features by stochastically filling the missing information. The main motivation behind ZS-SBIR lies with sketches being costly and labour-intensive to source – sketches need to be individually drawn by hand, other than crawled for free from the internet. To enable rapid deployment on categories where training sketches are not readily available, it is important to leverage on existing sketch data from other categories. The key difference between ZS-SBIR and other ZS tasks, which is also the main difficulty of the problem, lies with the additional modality gap between sketch and photo.

3. QuickDraw-Extended Dataset

Existing datasets do not cover all the challenges derived from a ZS-SBIR system. Therefore, we propose a new dataset named *QuickDraw-Extended Dataset* that is specially designed for this task. First we review the existing datasets in the literature used for ZS-SBIR and motivate the purpose of the new dataset. Thus, we provide a large-scale ZS-SBIR dataset that overcomes the main problems of the existing ones. Existing datasets were not originally designed for a ZS-SBIR scenario, but they have been adapted by a redefining the partitions setup. In addition, the main limitations that we overcome with the new dataset are (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval.

Sketchy-Extended Dataset [27]: Originally created as a fine-grained association between sketches to particular photos for fine-grained retrieval. This dataset has been adapted to the task of ZS-SBIR. On one hand, Shen *et al.* [28] proposed to set aside 25 random classes as a test set whereas the training is performed in the rest 100 classes. On the other hand, Yelamarthi *et al.* [36] proposed a different partition of 104 train classes and 21 test classes in order to make sure that test is not present in the 1,000 classes of ImageNet. Its main limitation for the task of ZS-SBIR is its fine-grained nature, i.e., each sketch has a corresponding photo that was used as reference at drawing time. Thus, participants tended to draw the objects in a realistic fashion, producing sketches resembling that of a true edge-map very well. This essentially narrows the cross-domain gap between sketch and photo.

TUBerlin-Extended Dataset [6]: It is a dataset that was created for sketch classification and recognition benchmarking. In this case, drawers were asked to draw the sketches giving them only the name of the class. This allows a semantic connection among sketches and avoids possible biases. However, the number of sketches is scarce, considering the variability among the observations of a concept in the real world. Also, some of the design decisions on the selection of object categories prevent it to be adequate for our zero-shot setting: (i) classes are defined both in terms of a concept and an attribute (e.g., seagull, flying-bird); (ii) different WordNet levels are used, *i.e.* there are classes that are semantically included in others (e.g., mug, beer-mug).

3.1. The Dataset

Taking into account the limitations of the previously described datasets in a ZS-SBIR scenario, we contribute to the community a novel large-scale dataset, *QuickDraw-Extended*. We identified the following challenges of a practical ZS-SBIR, (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval. According to this, the new dataset

Table 1. Dataset comparison in terms of their size. Partition is presented in terms of number of classes used for each set, moreover, # Comparisons stands for the number of comparisons sketch-image performed in test.

	Sketchy [27]	TUBerlin [6]	QuickDraw
Partition (tr+va, te)	(104, 21)	(220, 30)	(80, 30)
# Sketch/class	500	80	3,000
# Image/class	600-700	$\sim 764^a$	$\sim 1,854$
# Comparisons	$\sim 10\text{Mill.}$	$\sim 1.9\text{Mill.}$	$\sim 166\text{Mill.}$

^aExtremely imbalanced

must fulfil the following aspects: (i) to not have a direct one-to-one correspondence between sketches and images, *i.e.* sketches can be rough conceptual abstractions of images produced in an amateur drawing style; (ii) to avoid ambiguities and overlapping classes; (iii) large intra-class variability provided by the high abstraction level of different drawers.

In order to accomplish these objectives, we took advantage of the Google Quick, Draw! [12] data which is a huge collection of drawings (50 millions) belonging to 345 categories obtained from the *Quick, Draw!*[†] game. In this game, the user is asked to draw a sketch of a given category while the computer tries to classify them. The way sketches are collected provides the dataset a large variability, derived from human abstraction. Moreover, it addresses the large domain gap between non-expert drawers and photos that is not considered in previous benchmarks. Hence, we propose to make use of a subset of sketches to construct a novel dataset for large-scale ZS-SBIR containing 110 categories (80 for training and 30 for testing). Classes such as `circle` or `zigzag` are directly discarded because they can not be used in an appropriate SBIR. As a retrieval gallery, we provide images extracted from *Flickr* tagged with the corresponding label. Manual filtering is performed to remove outliers. Moreover, following the idea introduced in [36] for the *Sketchy-Extended* dateset, we provide a test split which forces that test classes are not present in ImageNet in case of using pre-trained models. Finally, this dataset consists of 330,000 sketches and 204,000 photos moving towards a large-scale retrieval. We consider that this dataset will provide better insights about the real performance of ZS-SBIR in a real scenario.

Table 1 provides a comparison of the three benchmarks for the task of ZS-SBIR. To the best of our knowledge, this is the first time that a real large-scale problem is addressed providing 6 times more sketches and more than the double of photos per each class. Qualitatively QuickDraw-Extended provides a high abstraction level than previous

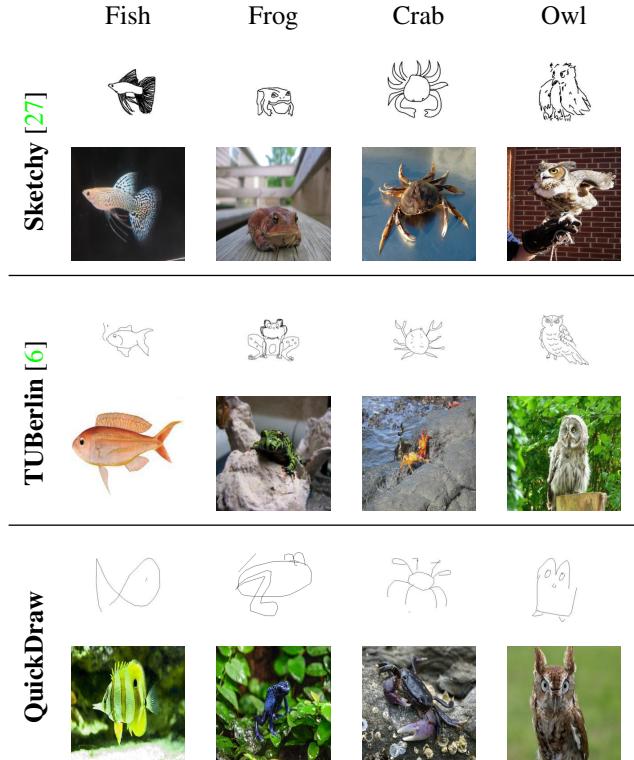


Figure 2. Qualitative comparison of the datasets. The different levels of abstraction in the sketches can be appreciated. From the top to the bottom, the figure also shows the decrease in the alignment between sketches and images.

benchmarks as it is shown in Figure 2.

4. A ZS-SBIR framework

4.1. Problem Formulation

Let \mathcal{C} be the set of all possible categories in a given dataset; $\mathcal{X} = \{x_i\}_{i=1}^N$ and $\mathcal{Y} = \{y_i\}_{i=1}^M$ be the set of photos and sketches respectively; $l_x : \mathcal{X} \rightarrow \mathcal{C}$ and $l_y : \mathcal{Y} \rightarrow \mathcal{C}$ be two labelling functions for photos and sketches respectively. Such that give an input sketch an optimal ranking of gallery images can be obtained. In a *zero-shot* framework, training and testing sets are divided according to *seen* $C^s \subset \mathcal{C}$ and *unseen* $C^u \subset \mathcal{C}$ categories, where $\mathcal{C}^s \cap \mathcal{C}^u = \emptyset$. Thus, the model needs to learn an aligned space between sketches and photos to perform well on test data whose classes have never been used in training. We define the set of *seen* and *unseen* photos as $\mathcal{X}^s = \{x_i ; l_x(x_i) \in C^s\}_{i=1}^N$ and $\mathcal{X}^u = \mathcal{X} \setminus \mathcal{X}^s$. We define analogously the *seen* and *unseen* sets for sketches, denoted as \mathcal{Y}^s and \mathcal{Y}^u .

The proposed framework is divided in two main components. The encoder transforms the input image to the corresponding embedding space. The second component is the cost function which guides the learning process to provide

[†]<https://quickdraw.withgoogle.com/>

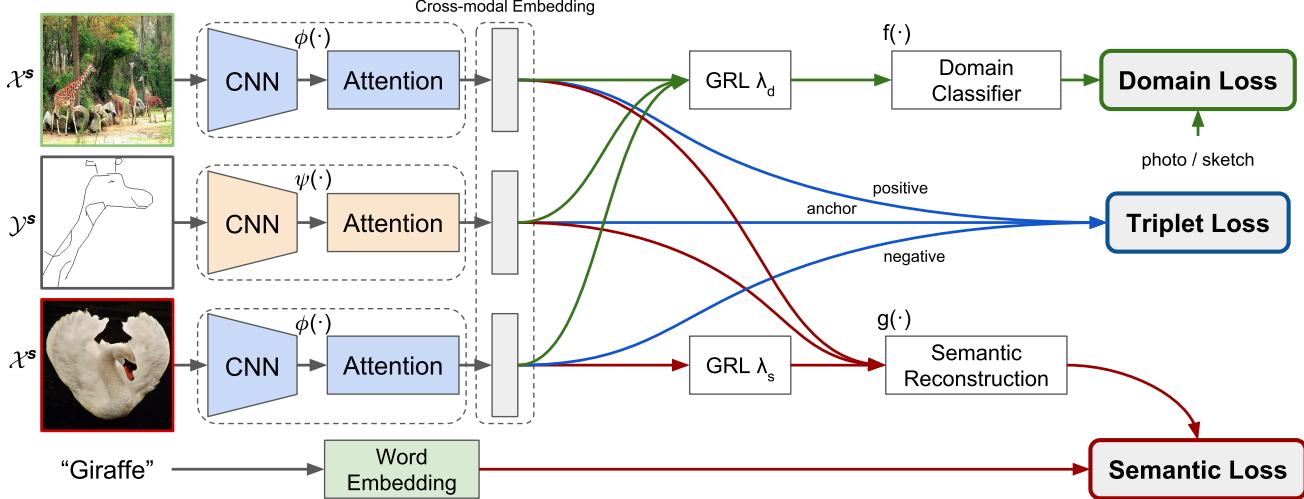


Figure 3. Proposed architecture for ZS-SBIR which maps sketches and photos in a common embedding space. It combines three losses: (i) triplet loss, to learn a ranking metric; (ii) domain loss to merge images and sketches to an indistinguishable space making use of a GRL; (iii) semantic loss forces the embeddings to contain semantic information by reconstructing the word2vec embedding of the class. It also helps to distinguish semantically similar classes by means of a GRL on the negative example (best viewed in color).

the embedding with the desired properties. Figure 3 outlines the proposed approach.

4.2. Encoder Networks

Given a distance function $d(\cdot, \cdot)$, the aim of our framework is to learn two embedding functions $\phi : \mathcal{X} \rightarrow \mathbb{R}^D$ and $\psi : \mathcal{Y} \rightarrow \mathbb{R}^D$ which respectively map the photo and sketch domain into a common embedding space. Later, these embedding functions are used in the retrieval task during the test phase, therefore, they should possess a ranking property related to the considered distance function. Hence, given two photos $x_1, x_2 \in \mathcal{X}$ and a sketch $y \in \mathcal{Y}$, we expect the embedding fulfills the following condition: $d(\phi(x_1), \psi(y)) < d(\phi(x_2), \psi(y))$, when $l_x(x_1) = l_y(y)$ and $l_x(x_2) \neq l_y(y)$. In a retrieval scenario, our system is able to provide a ranked list of images by the chosen distance function. In this framework, d has been set as ℓ_2 -distance. During training, the two embedding $\phi(\cdot)$ and $\psi(\cdot)$ are trained with multi-modal information, therefore they presume to learn a modality free representation.

Our embedding functions $\phi(\cdot)$ and $\psi(\cdot)$ are defined as two CNNs with attention where the last fully-connected layer has been replaced to match the desired embedding size D . The *attention* [33] mechanism helps our system to localise the important features in both modalities. Soft-attention is the widely used one because it is differentiable, and hence it can be learned end-to-end with the rest of the network. Our soft-attention model learns an attention mask which assigns different weights to different regions of an image given a feature map. These weights are used to highlight important features, therefore, given an attention mask

att and a feature map f , the output of the attention module is computed by $f + f \cdot att$. The attention mask is computed by means of 1×1 convolution layers applied on the corresponding feature map.

4.3. Learning objectives

The learning objective of the proposed framework combines: (i) *Triplet Loss*; (ii) *Domain Loss*, (iii) *Semantic Loss*. These objective functions provide visual and semantic information to the encoder network. Let us consider a triplet $\{a, p, n\}$ where $a \in \mathcal{Y}^s$, $p \in \mathcal{X}^s$ and $n \in \mathcal{X}^s$ are respectively the *anchor*, *positive* and *negative* samples during the training. Moreover, $l_x(p) = l_y(a)$ and $l_x(n) \neq l_y(a)$.

Triplet Loss: This loss aims to reduce the distance between embedded sketch and image if they belong to the same class and increase it if they belong to different classes. For simplicity, if we define the distances between the samples as $\delta_+ = \|\psi(a) - \phi(p)\|_2$ and $\delta_- = \|\psi(a) - \phi(n)\|_2$ for the *positive* and *negative* samples respectively, then, the ranking loss for a particular triplet can be formulated as $\lambda(\delta_+, \delta_-) = \max\{0, \mu + \delta_+ - \delta_-\}$ where $\mu > 0$ is a margin parameter. Batch-wise, the loss is defined as:

$$\mathcal{L}_t = \frac{1}{N} \sum_{i=1}^N \lambda(\delta_+^i, \delta_-^i). \quad (1)$$

This loss measures the violation of the ranking order of the embedded features. Therefore, the order aimed by this loss is $\delta_- > \delta_+ + \mu$, if this is the case, the network is not updated, otherwise, the weights of the network are updated accordingly. Triplet loss provides a metric space with ranking properties based on visual features.

Domain Loss: Triplet loss mentioned above does not explicitly enforce the mapping of sketch and image samples to a common space. Therefore, at this end, to ensure that the obtained embedding belong to the same space, we propose to use a domain adaptation loss [8]. The basic idea of this loss is to obtain a domain-agnostic embedding that does not contain enough information to decide whether it comes from a sketch or photo. Given the embedding $\phi(\cdot)$ and $\psi(\cdot)$, we make use of a Multilayer Perceptron (MLP) as a binary classifier trying to predict which was the initial domain. Purposefully, in order to create indistinguishable embedding we use a *GRL* defined as $R_\lambda(\cdot)$, which applies the identity function during the forward pass $R_\lambda(x) = x$, whereas during the backward pass it multiplies the gradients by the meta-parameter $-\lambda$, $\frac{dR_\lambda}{dx} = -\lambda I$. This operation reverses the sign of the gradient that flows through the CNNs. In this way, we encourage our encoders to extract the shared representation from sketch and photo. For this loss, we define a meta-parameter λ_d that changes from 0 (only trains the classifier but does not update the encoder network) to 1 during the training according to a defined function. In our case it is defined according to the iteration i as $z_\lambda(i) = (i - 5)/20$. Following the notation, $f : \mathbb{R}^D \rightarrow [0, 1]$ be the MLP and $e \in \mathbb{R}^D$ an embedding coming from the encoders network. Then we can define the binary cross entropy of one of the samples as $l_t(e) = t \log(f(R_{\lambda_d}(e))) + (1 - t) \log(1 - f(R_{\lambda_d}(e)))$, where e is the embedding obtained by the encoder network and t is 0 and 1 for sketch and photo domains respectively. Hence, the domain loss is defined as:

$$\mathcal{L}_d = \frac{1}{3N} \sum_{i=1}^N (l_0(\psi(a_i)) + l_1(\phi(p_i)) + l_1(\phi(n_i))) \quad (2)$$

Semantic Loss: A decoder network trying to reconstruct the semantic information of the corresponding category from the generated embedding is proposed. This reconstruction forces that the semantic information is encoded in the obtained embedding. In this case, we propose to minimise the cosine distance with the reconstructed feature vector and the semantic representation of the category. Inspired by the idea presented by Gonzalez *et al.* [9] for cross-domain disentanglement, we propose to exploit the negative sample to foster the difference between similar semantic categories. Hence, we apply a GRL $R_{\lambda_s}(\cdot)$ to the negative sample at the input of the semantic decoder and we train it to reconstruct the semantics of the positive example. The idea is to help the encoder network to separate the semantic information of similar classes. In this case, we decided to keep the meta-parameter λ_s to a fixed value among all the training, in particular, it was set to 0.5.

Let $c \in \mathcal{C}^s$ be the corresponding category of the anchor a . The semantics of this category are obtained by the

word2vec [20] embedding trained on part of Google News dataset (~ 100 billion words), *GloVe* [23] and *fastText* [1] (more results are available in supplementary materials). Let $g : \mathbb{R}^D \rightarrow \mathbb{R}^{300}$ be the semantic reconstruction network and $s = \text{embedding}(c) \in \mathbb{R}^{300}$ be the semantics of the given category. Hence, given an image embedding $e \in \mathbb{R}^D$ the cosine loss is defined as $l_c(e, s) = \frac{1}{2} \left(1 - \frac{g(e)s^t}{\|g(e)\|\|s\|} \right)$. The semantic loss is defined as follows:

$$\begin{aligned} \mathcal{L}_s = & \frac{1}{3N} \sum_{i=1}^N (l_c(\psi(a_i), s_i) + l_c(\phi(p_i), s_i) \\ & + l_c(R_{\lambda_s}(\phi(n_i)), s_i)) \end{aligned} \quad (3)$$

Therefore, the whole network will be trained by a combination of three proposed loss functions.

$$\mathcal{L} = \alpha_1 \mathcal{L}_t + \alpha_2 \mathcal{L}_d + \alpha_3 \mathcal{L}_s, \quad (4)$$

where the weighting factors α_1 , α_2 and α_3 are equal in our model. Algorithm 1 presents the training algorithm followed in this work. $\Gamma(\cdot)$ denotes the optimiser function.

Algorithm 1 Training algorithm for the proposed model .

Input: Photo-Sketch data $\{\mathcal{X}, \mathcal{Y}\}$; Class semantics \mathcal{S} ;
 $\lambda_s = 0.5$ and max training iterations T
Output: Encoder networks parameters $\{\Theta_\phi, \Theta_\psi\}$.

```

1: repeat
2:   Get a random mini-batch  $\{y_i, x_i^p, x_i^n, s_i\}_{i=1}^{N_B}$ ; where
3:      $y_i, x_i^p$  belong to the same class and  $x_i^n$  does not.
4:    $\lambda_d \leftarrow \text{clip}(z_\lambda(\cdot), \min = 0, \max = 1)$ 
5:    $\mathcal{L} \leftarrow \text{Eq. 4}$ 
6:    $\Theta \leftarrow \Theta - \Gamma(\nabla_\Theta \mathcal{L})$ 
7: until Convergence or max training iterations  $T$ 
```

5. Experimental Validation

This Section experimentally validates the proposed ZS-SBIR approach on three benchmarks *Sketchy-Extended*, *TUBerlin-Extended* and *QuickDraw-Extended*, highlighting the importance of the newly introduced dataset which is more realistic for practical SBIR purpose. A detailed comparison with the state-of-the-art is also presented.

5.1. Zero-shot Experimental Setting

Implementation details: Our CNN-based encoder networks $\phi(\cdot)$ and $\psi(\cdot)$ make use of a ImageNet pre-trained VGG-16 [29] architecture. This can be replaced by any model to enhance the extracted feature quality. Both, domain classifier $f(\cdot)$ and semantic reconstruction $g(\cdot)$ of the proposed model makes use of 3 fully connected layers with ReLU activation functions. The whole framework was implemented with PyTorch [22] deep learning tool and is trainable on single Pascal Titan X GPU card.

Table 2. Comparison against the state-of-the-art with that of the proposed model. Note: the same train and test split are used for all experiments on CVAE [36] and ours. ZSIH [28] did not report the specific details on their split (other than 25 classes were used for testing), and we could not produce their results on *QuickDraw-Extended* due to the lack of publicly available code.

Method	Sketchy-Extended [27]			TUBerlin-Extended [6]			QuickDraw-Extended		
	mAP	mAP@200	P@200	mAP	mAP@200	P@200	mAP	mAP@200	P@200
ZSIH [28]	0.2540 ^a	—	—	0.2200	—	—	Not able to produce		
CVAE [36]	0.1959	0.2250	0.3330	0.0050	0.0090	0.0030	0.0030	0.0060	0.0030
Ours	0.3691	0.4606	0.3704	0.1094	0.1568	0.1208	0.0752	0.0901	0.0675

^aUsing a random partition of 25 test categories following the setting proposed in [26], we obtained 0.3521 for our model.

Table 3. Ablation study for the proposed model. As baseline, the triplet loss is used and the different modules are incrementally added.

Attn. Dom. Sem.	Sketchy-Extended [27]			TUBerlin-Extended [6]			QuickDraw-Extended		
	mAP	mAP@200	P@200	mAP	mAP@200	P@200	mAP	mAP@200	P@200
- - -	0.3020	0.3890	0.3091	0.0590	0.1040	0.0682	0.0354	0.0546	0.0454
✓ - -	0.3207	0.4150	0.3342	0.0729	0.1141	0.1002	0.0456	0.0635	0.0496
✓ ✓ -	0.3256	0.4113	0.3444	0.0845	0.1264	0.1080	0.0651	0.0881	0.0615
✓ - ✓	0.3392	0.4146	0.3586	0.1055	0.1496	0.1115	0.0693	0.0896	0.0625
✓ ✓ ✓	0.3691	0.4606	0.3704	0.1094	0.1568	0.1208	0.0752	0.0901	0.0675

Training setting: Our system uses triplets to utilise the inherent ranking order. The training batches are constructed in a way so that it can take the advantage of the semantic information in order to mine hard negative samples for a given anchor class. This implies that semantically closer classes will have a higher probability to be used during training and thus they are likely to be disjoint in the final embedding. We trained our model following an early stopping strategy in validation to provide the final test result. The model is trained end-to-end using the SGD [2] optimiser. The learning rate used throughout is $1e - 4$. The epochs required to train the model on different dataset is around 40.

Evaluation protocol: The proposed evaluation uses the metrics used by Yelamarthi *et al.* [36]. Therefore, the evaluation is performed taking into account the top 200 retrieved samples. Moreover, we also provide metrics on the whole dataset. Images labelled with the same category as that of the query sketch, are considered as relevant. Note that this evaluation does not consider visually similar drawings that can be considered correct by human users. For the existing datasets, we used the proposed splits in [36, 28].

5.2. Model Discussion

This section presents a comparative study with the state-of-the-art followed by a discussion on the *TUBerlin-Extended* results and finally the ablative study. As mentioned, our model is build on top of a triplet network. We take this as a baseline and study the importance of the different components of the full model which includes the *attention mechanism*, the *semantic loss* and the *domain loss*.

Comparison: Table 2 provides comparisons of our full model results against those of the state-of-the-art. We report a comparative study with regard to two methods presented in Section 2, namely ZSIH [28] and CVAE [36]. Note that we have not been able to reproduce the ZSIH model due to lack of technical implementation details and the code being unavailable. Hence, the results on *QuickDraw-Extended* dataset nor an evaluation using the top 200 retrieval could be computed. The last row of the Table 2 shows the result of our full model. From the Table 2 the results suggest the limitation of the previous models regarding their ability in an unconstrained domain where sketches have higher level of abstractions. The CVAE [36] method trained with sketch-image correspondence has difficulties to capture the intra-class variability, the domain gap and also the ability to infer unseen classes. The following conclusions are drawn: (i) our base model outperforms all the state-of-the-art methods in *Sketchy-Extended* Dataset; (ii) our model performs the best overall on each metric and on almost all the datasets; (iii) the gap between our model and the state-of-the-art datasets is almost double in *Sketchy-Extended* Dataset; (iv) the difference in the result in previous dataset points out the need of a new well structured dataset for ZS-SBIR (v) the new benchmark also provides the different aspects (*i.e* of semantics, mutual information) that can play important role in a real ZS-SBIR scenario; (vi) the evaluation shows the importance of going towards large-scale ZS-SBIR where the retrieval search space is in the range of 166 million comparisons (16 times of the current largest dataset).

Discussion on TUBerlin-Extended: As stated in Section 3,

Sketchy [27]										QuickDraw									
Query	Top-8 retrieved candidates								Query	Top-8 retrieved candidates									
CVAE [36]																			
Ours pear																			
CVAE [36]																			
Ours door																			

Figure 4. Top 8 image retrieval examples given a query sketch. All the examples correspond to a zero-shot setting, *i.e.* no example have been seen in training. First row provides a comparison with CVAE [36] method against our pipeline. Note that in some retrieval cases, for instance, *door* is confused with *window* images which can be true even for humans. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)

the results could be heavily affected by the chosen classes for experiments. Since [28] did not report specific details on their train and test split, we can not offer a fair comparison on TUBerlin-Extended. Instead, for both [36] and ours, we resort to the commonly accepted median over random splits setting. And it shows our method favourably beats [36] by a clear margin. We did however observe a high degree of fluctuation over the different splits on TUBerlin-Extended, which re-affirms our speculation on how the categories included in TUBerlin-Extended might not be optimal for the zero-shot setting (see Section 3). This could explain the superior performance of [28], yet more experiments are needed to confirm such suspicion. Unfortunately, again such experiments would not be possible without details on their train and test split.

Ablation study: Here, we investigate the contribution of each component to the model, as well as other issues of the architecture. The first 5 rows of Table 3 present a study of the contribution of each component to the whole proposed model. From this Table we can draw the following conclusions: (i) attention plays a major role in improving the baseline result; (ii) the domain loss is able to alleviate to some extend the domain gap, this is more remarkable in those datasets where sketches are more abstract; (iii) as the difficulty of the dataset increases, the semantic and the domain losses start playing a major role in improving the baseline result; (iv) semantics provide better extrapolation to unseen data than domain loss which shows that either the mutual information is very less or that the semantic information is really needed in this extrapolation; (v) the poor performance in the *QuickDraw-Extended* dataset shows that the practical problem of ZS-SBIR is still indeed unsolved. It should be noted, that the best model makes use of the three losses.

Qualitative: Some retrieval results are shown in Figure 4 for *Sketchy-Extended* and *QuickDraw-Extended*. We also

provide a qualitative comparison with CVAE proposed by Yelamarthi *et al.* [36]. The qualitative results reinforce that the combination of semantic, domain and triplet loss fairs well in a dataset with substantial variances on visual abstraction. We would also like to point out that the retrieved results for the class *skyscraper* show high visual shape similarity with rectangle *i.e.* *door* and *saw*. The retrieved circular *saw* could also might be retrieved because of the semantic rather than the visual similarity. Similar visual correspondences can also be noticed between the query sketch *helicopter* and the retrieved result *windmill*.

6. Conclusions

This paper represents a first step towards a practical ZS-SBIR task. Previous works on this task do not address some of the important challenges that appear when moving to an unconstrained retrieval and do not tackle with the large domain gap between amateur sketch and photo. In this scenario, to overcome the lack of proper data, we have contributed to the community a specifically designed large-scale ZS-SBIR dataset, *QuickDraw-Extended* which provides highly abstract amateur sketches collected with the Google Quick, Draw! game. Then, we have proposed a novel ZS-SBIR system that combines visual as well as semantic information to generate an image embedding. We experimentally show that this novel framework overcomes recent state-of-the-art methods in the ZS-SBIR setting.

Acknowledgements

Work supported by EU’s MSC grant No. 665919, Spanish grants FPU15/06264 and TIN2015-70924-C2-2-R; and the CERCA Program/Generalitat de Catalunya. The Titan X was donated by NVIDIA. This work was carried out during research stay at SketchX Lab in QMUL.

References

- [1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 2017. 6
- [2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*. 2010. 7
- [3] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, 2017. 3
- [4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2
- [5] Sounak Dey, Anjan Dutta, Suman Kumar Ghosh, Ernest Valveny, Josep Lladós, and Umapada Pal. Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In *ICPR*, 2018. 1
- [6] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? In *SIGGRAPH*, 2012. 1, 2, 3, 4, 7
- [7] Mathias Eitz, Kristian Hildebrand, Tammy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on VCG*, pages 1624–1636, 2011. 2
- [8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2, 6
- [9] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. *NeurIPS*, 2018. 6
- [10] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 117(7):790–806, 2013. 2, 3
- [11] Stuart James, Manuel J Fonseca, and John Collomosse. Reenact: Sketch based choreographic design from archival dance footage. In *ICMR*, 2014. 3
- [12] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. The quick, draw! - a.i. experiment. <https://quickdraw.withgoogle.com>, 2016. 2, 4
- [13] Nour Karessli, Zeynep Akata, Bernt Schiele, Andreas Bulling, et al. Gaze embeddings for zero-shot image classification. In *CVPR*, 2017. 3
- [14] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE PAMI*, 36(3):453–465, 2014. 3
- [15] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014. 3
- [16] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueling Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *CVPR*, 2017. 1
- [17] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 1, 2, 3
- [18] Yang Long, Li Liu, Yuming Shen, Ling Shao, and J Song. Towards affordable semantic searching: Zero-shot retrieval via dominant attributes. In *AAAI*, 2018. 3
- [19] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 2
- [20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 6
- [21] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 3
- [22] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch, 2017. 6
- [23] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 6
- [24] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, 2016. 1
- [25] Jose M Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *ICIP*, 2014. 2
- [26] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, 2015. 2
- [27] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *SIGGRAPH*, 2016. 1, 2, 3, 4, 7, 8
- [28] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 3, 7, 8
- [29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, abs/1409.1556, 2014. 6
- [30] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017. 1, 2, 3
- [31] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015. 3
- [32] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *ICCV*, 2017. 3
- [33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 5
- [34] Xing Xu, Fumin Shen, Yang Yang, Dongxiang Zhang, Heng Tao Shen, and Jingkuan Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *CVPR*, 2017. 3
- [35] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *CVPR*, 2017. 3
- [36] Sasikiran Yelamarthi, Shiva Krishna Reddy M, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 3, 4, 7, 8
- [37] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 2, 3

- [38] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016. [2](#), [3](#)
- [39] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, 2018. [3](#)
- [40] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. [3](#)