BDA-SketRet: Bi-Level Domain Adaptation for Zero-Shot SBIR

Ushasi Chaudhuri[a,*], Ruchika Chavan[b], Biplab Banerjee[a], Anjan Dutta[c], Zeynep Akata[d]

[a]*Indian Institute of Technology Bombay, India*
[b]*School of Informatics, The University of Edinburgh, United Kingdom*
[c]*Computer Science, University of Exeter, United Kingdom*
[d]*Computer Science, University of Tübingen, Germany*

Abstract

The efficacy of zero-shot sketch-based image retrieval (ZS-SBIR) models is governed by two challenges. The immense distributions-gap between the sketches and the images requires a proper domain alignment. Moreover, the fine-grained nature of the task and the high intra-class variance of many categories necessitates a class-wise discriminative mapping among the sketch, image, and the semantic spaces. Under this premise, we propose BDA-SketRet, a novel ZS-SBIR framework performing a bi-level domain adaptation for aligning the spatial and semantic features of the visual data pairs progressively. In order to highlight the shared features and reduce the effects of any sketch or image-specific artifacts, we propose a novel symmetric loss function based on the notion of information bottleneck for aligning the semantic features while a cross-entropy-based adversarial loss is introduced to align the spatial feature maps. Finally, our CNN-based model confirms the discriminativeness of the shared latent space through a novel topology-preserving semantic projection network. Experimental results on the extended Sketchy, TU-Berlin, and QuickDraw datasets exhibit sharp improvements over the literature.

Keywords: Sketch-based image retrieval, zero-shot learning, domain adaptation, generalized zero-shot learning, graph convolution.

1. Introduction

Sketches are the most minimalistic representation of visual data. With the advancement in sensor technology, a quick hand-drawn sketch can be used as a query image to retrieve similar category samples from the visual domain. This is especially useful when there is a lack of available visual query sample at hand, while it is only at the mind of the user at a vague picture (Xu et al., 2020a). In traditional sketch-based image retrieval (SBIR) (Eitz et al., 2010), the training and the test classes are the same. However, this is unrealistic as the model may encounter novel classes during inference in many on-the-fly retrieval applications. The *zero-shot learning* (ZSL) (Xian et al., 2017, 2018; Romera-Paredes and Torr, 2015) aims to bridge the gap between the non-overlapping sets of training classes (*seen*) and test classes (*unseen*) using semantic side information for visual recognition. To this end, attempts have been made to integrate ZSL with SBIR to get the zero-shot sketch-based image retrieval (ZS-SBIR) problem (Yelamarthi et al., 2018), even without constraining the search space to contain only the unseen classes at test time, e.g., generalized ZS-SBIR (GZS-SBIR) (Dutta and Akata, 2020). This essentially requires aligning the visual and semantic features (which is used for describing the class) in the embedding space. In this paper, we aim to solve the (G)ZS-SBIR problem by looking into some of the existing problems that are explained in the following subsections and propose improved solutions to this end.
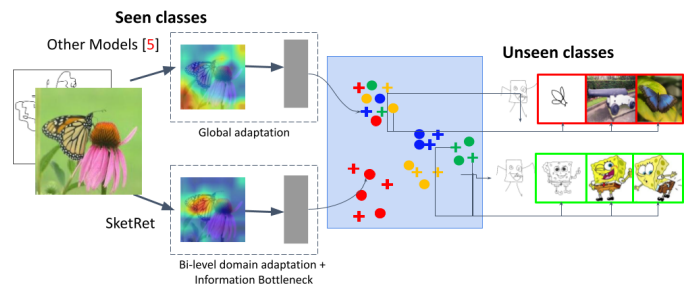

Figure 1: Given a cluttered feature space for images and sketches, a sub-optimal domain alignment fails to produce a discriminative latent space which is affected by problems like hubness and negative transfer. Feature adaptation at multiple feature scales together with a discriminative feature space learning ensures that zero-shot testing can be performed well in BDA-SketRet.

1.1. Observations

A key challenge in solving ZS-SBIR is because of the fact that sketches are usually drawn by various artists. ZS-SBIR (Yelamarthi et al., 2018; Dutta and Akata, 2020) models need to overcome a substantial within-category variance in addition to the domain gap between sketches and images given their disparity in spectral, spatial, and texture properties (Xian et al., 2017, 2018; Romera-Paredes and Torr, 2015). Moreover, natural images have random background effects which are completely absent in the sketches. In this regard, the majority of the existing ZS-SBIR approaches learn independent mappings to align the visual domains either in a semantic space (Dutta and Akata, 2019; Pandey et al., 2020) or in a semantics influ-

---

*Corresponding author
Email address: ushasi@iitb.ac.in (Ushasi Chaudhuri)

enced latent space (Dey et al., 2019). In both cases, the representations are extracted from the entire image / sketch data while largely neglecting the complex distributions of different local regions. This results in sub-optimal domain alignment as not all the information is equally transferable. Another challenge is that the performance on the unseen classes in generalized (G)ZS-SBIR is often compromised since the model is biased towards the seen classes. A discriminative visual-semantic mapping is generally preferred to alleviate this issue.

Further, images and sketches that are significantly dissimilar across domains in the feature space should not be forcefully aligned; otherwise, the model becomes vulnerable to the negative transfer of irrelevant knowledge. This dilemma can be curtailed if the global feature extraction is guided by highly domain-independent mid-level regional properties. The existing work in the ZS-SBIR literature that use domain adaptation primarily tries to align the final features of the sketch and images (Dutta and Akata, 2019; Dey et al., 2019; Dutta et al., 2020). This often leads to a sub-optimal domain alignment due to variations in the sketch domain. These methods are non-trivial to use in the ZS-SBIR setup given the disjoint nature of the training and test classes. As the quality of the spatial feature maps obtained from the intermediate layers affects global image-level features extracted from deeper layers in a CNN, it becomes hard to forcefully align the global features without ensuring equivalence of the preceding feature maps. Apparently, the feature maps look substantially different for the sketch and the image modalities considering several issues like the variations in the sketch drawings and background effects in case of images. We argue that the well-aligned mid-level feature maps would supplement the global feature adaptation process. Hence, we introduce a ZS-SBIR model called BDA-SketRet which inculcates a rigorous multi-level modality adaptation policy between sketches and images both at the mid and high level feature embeddings to ensure a highly domain-agnostic, discriminative, and task-specific representation learning. Fig. 1 illustrates in briefs the main idea behind the proposed BDA-SketRet.

1.2. Our Approach

In BDA-SketRet, we propose to apply cross-modal visual adaptation separately at the local and global levels. The local adaptation is essential as the mid-level feature maps are very dissimilar between the domains and hence making them indistinguishable at this level is challenging. The major problem in aligning the feature maps arises from the vagueness of the decision boundary between the domains. Hence, we propose an cross-entropy based loss measure where the ground truth probability is selected to have high entropy. In the global adaptation, we aim to make the high-level feature embeddings more discriminative and devoid of any irrelevant domain-specific information which may hamper the alignment. For this, we introduce a novel metric learning based information theoretic alignment loss based on symmetric KL divergence. Finally, since the sketch and image modalities are vastly different, we include two cross-modal mid-to-high level feature reconstruction modules which further boosts the domain invariance of the backbone networks. This is different from the traditional cross-reconstruction of the high-level features typically followed in the multi-modal learning tasks. On the other hand, we propose a semantic projection network for the class prototypes which preserves the neighborhood topology of the original semantic space into the embedding space. Generally, the ZSL models are biased towards to training classes which may partially be solved with neighborhood preservation. In order to accomplish the same, we introduce a network consisting of MLP and graph CNN (GCN).

Our major contributions are as follows: i) We introduce BDA-SketRet, a discriminative multi-level domain adaptation framework for solving ZS-SBIR and GZS-SBIR. ii) We showcase the effectiveness of aligning both the domains in two different feature levels (mid and high level) and propose a novel strategy for the same. A novel symmetric KL divergence based formulation is introduced for information bottleneck loss to be used for semantic feature alignment as opposed to the traditional asymmetric KL based VIB loss measure (Tishby and Zaslavsky, 2015). iii) An intuitive neighborhood-preserving semantic projection module is proposed which is seen to reduce the model bias towards the seen classes specifically for GZS-SBIR. iv) We conduct extensive experiments on the extended Sketchy, TU-Berlin, and QuickDraw datasets and rigorously ablate the model. The efficacy of the proposed model is verified experimentally as the proposed network beats the existing state-of-the-art models by a margin of $\approx$ 3 - 4% in all the evaluation metrices.

2. Related Works

In this section, we review prior related works on ZS-SBIR, semantic projection and GCN that are related to ours. We also show how is our work different from the existing state-of-the-art models.

2.1. State-of-the-Art ZS-SBIR Algorithms

Prior to solving ZS-SBIR tasks, the simple SBIR problem received a lot of attention (Lei et al., 2019). As already stated, the major obstacle in solving the SBIR task stems from the fact that the distributions difference between sketch and image data is exceedingly large. Early works in this area include conventional pattern recognition methods for retrieval by engineering hand-crafted visual features (Hu and Collomosse, 2013; Saavedra, 2014). The proposition behind such approaches is to solve the task at hand by obtaining the edge-map of the natural images and to further match them with sketches arising from the same categories. As expected, the low-level SIFT (Lowe, 1999), SURF (Bay et al., 2006), or HoG (Dalal and Triggs, 2005) based descriptors are unable to properly encode the regional variations of the sketch data, resulting in an inferior cross-modal matching. The performance measures of deep CNN based SBIR models have witnessed a massive enhancement lately, thanks to the data-driven feature learning capabilities of CNN. Since the retrieval performance benefits from a discriminative feature space, several endeavors rely on distance

metric learning strategies like contrastive-loss (Chopra et al., 2005), triplet-loss (Sangkloy et al., 2016), and HOLEF-based loss (Song et al., 2017), to name a few. (Bui et al., 2017) proposed an efficient representation for sketch and image features using triplet loss, while on the other hand (Federici et al., 2020) provides an information-theoretic approach. Likewise, (Song et al., 2017; Qi et al., 2016; Yu et al., 2017; Sangkloy et al., 2016; Wang et al., 2015; Liu et al., 2017; Zhang et al., 2018) aims to solve a similar task.

The ZS-SBIR literature consists of both the discriminative and generative deep learning based techniques. Under the generative umbrella, (Shen et al., 2018) proposes a hashing network for the semantic knowledge reconstruction (ZSIH). Similarly, (Yelamarthi et al., 2018) introduces a conditional generative model for ZS-SBIR based on variational learning. The stacked auto-encoder (SAN) method (Pandey et al., 2020) deploys a generative framework based on stacked-adversarial networks within a Siamese architecture. The paired cyclic consistency loss proposed in SEM-PCYC (Dutta and Akata, 2019, 2020) helps in aligning the sketches and images in an encoded semantic space using adversarial training. Inspired by style transfer, (Dutta and Biswas, 2019; Dutta et al., 2020) develops a style-guided image to image translation model for ZS-SBIR, while (Dey et al., 2019) uses a triplet-based network to solve the task at hand. (Dutta et al., 2020) highlights the implications of data and class imbalance in ZS-SBIR and introduces an adaptive margin diversity regularizer (AMD-reg) to combat the same. As opposed to the real-valued feature embedding, hash-code based representations are also considered in this regard which offers a trade-off between performance and storage (Liu et al., 2017). The generative models for cross-modal style transfer are also explored (Dutta and Biswas, 2019) in this regard. While all the techniques showcase their performance on ZS-SBIR, a few works (Dutta and Akata, 2019, 2020; Pandey et al., 2020) also demonstrate their experiments for the GZS-SBIR setting.

2.2. Semantic projection and GCN.

The semantic information has an important role to play for attaining an improved ZS-SBIR inference. Generally, the visual modalities are aligned to a fixed semantic space. While (Dutta and Akata, 2019) uses an encoded semantic space obtained via a semantic auto-encoder, methods of (Dey et al., 2019; Dutta and Biswas, 2019) reconstruct the original semantic vectors from the visual information. Generating visual information from semantic prototypes has been dealt with in this regard (Yelamarthi et al., 2018). The graph CNN (Kipf and Welling, 2016) can encode the structural similarity among several objects and have been successfully applied for vision tasks like ZSL, image captioning, and visual question answering (Kampffmeyer et al., 2019; Yang et al., 2019; Narasimhan et al., 2018). The graph CNN has been considered in ZS-SBIR recently (Zhang et al., 2020) where the modality alignment takes place by restricting the domain-specific graphs to be isomorphic.

How are we different? Contrary to the existing literature, we choose to look into the ZS-SBIR problem mainly as a domain adaptation task wherein we propose to do a multi-level adaptation by tackling both spatial and semantic domain drifts. The notion of multi-level domain adaptation is new in this regard, to the best of our knowledge. From the theoretical point of view, we define a new symmetric loss function for the information bottleneck principle which is found to be well-suited for multi-modal labeled data pairs, something that the existing asymmetric divergence based functions fail to model. Finally, our semantic projection network offers topology preservation within minimal overhead and combats hubness effectively where all the existing techniques overlook this aspect completely.

3. Problem definition and preliminaries

Let $\mathcal{Z}^s = \{\mathcal{A}^s, \mathcal{B}^s, C^s, \mathcal{W}^s\}$ be a multi-modal training dataset consisting of images $\mathcal{A}^s$ and sketches $\mathcal{B}^s$ obtained from the $|C^s|$ *seen* visual categories. Additionally, we have access to semantic side information $\mathcal{W}^s$ which typically corresponds to the distributed word-vector embeddings of the individual category names. During inference, image and sketch data $\mathcal{Z}^u = \{\mathcal{A}^u, \mathcal{B}^u\}$ from a non-overlapping set of previously *unseen* classes C^u are considered ($C^u \cap C^s = \varnothing$) in the zero shot SBIR setup. We deal with the unpaired dataset setting in $\mathcal{Z}^s$ where the number of sketch and image instances in $\mathcal{A}^s$ and $\mathcal{B}^s$ are different: $\{a_i^s\}_{i=1}^N \in \mathcal{A}^s$ and $\{b_i^s\}_{i=1}^M \in \mathcal{B}^s$. The model is trained to reduce the distribution mismatch between $\mathcal{A}^s$ and $\mathcal{B}^s$ and subsequently to transfer the knowledge from $\mathcal{Z}^s$ to $\mathcal{Z}^u$ with the help of the semantic information $\mathcal{W}^s$. The testing phase concerns the retrieval of images with similar semantic categories from $\mathcal{A}^u$ given the sketch queries from $\mathcal{B}^u$. In contrast to ZS-SBIR, GZS-SBIR assumes the presence of images from $\mathcal{A}^s \cup \mathcal{A}^u$ during testing for unseen-class sketch queries coming from $\mathcal{B}^u$.

4. BDA-SketRet

The goal of BDA-SketRet is to align the images and sketches from the same class in a semantically meaningful shared latent space. It is composed of cross-modal triplets where the sketch data from $\mathcal{B}^s$ serves as the anchor (α) while the positive (p) and negative (n) counterparts are selected from $\mathcal{A}^s$ (Fig. 2). The feature networks for $\mathcal{A}^s$ and $\mathcal{B}^s$ are defined by $\phi(\cdot)$ and $\psi(\cdot)$ which are convolutional neural networks with integrated attention sub-networks $\phi_{\mathrm{att}}(\cdot)$ and $\psi_{\mathrm{att}}(\cdot)$. The attention block outputs are simultaneously projected to the local adversarial domain classifier $l(\cdot)$ to highlight spatially indistinct features of the same-class samples from $\mathcal{A}^s$ and $\mathcal{B}^s$ and to the shared latent space. We further introduce two cross-modal feature reconstruction modules $(\mathcal{V}_\alpha(\cdot), \mathcal{V}_p(\cdot))$ which aim to reconstruct $\phi(p)$ from $\psi_{\mathrm{att}}(\alpha)$ and $\psi(\alpha)$ from $\phi_{\mathrm{att}}(p)$ through variational bottlenecks. On the other hand, the outputs of $\phi(\cdot)$ and $\psi(\cdot)$ need to be synchronized for defining the shared embedding space. In this regard, the global domain adaptation on $\phi(p/n)$ and $\psi(\alpha)$ is carried out considering a combination of the domain classifier $f(\cdot)$ and a multi-class category classifier $h(\cdot)$. A semantic sub-network $g(\cdot, \cdot)$ comprising of an MLP $g_1(\cdot)$ and a graph CNN $g_2(\cdot, \cdot)$ is used to non-linearly project the semantic vectors into the shared space. The outputs of $g_1(\cdot)$ and $g_2(\cdot, \cdot)$ are concatenated and projected onto the latent space by another MLP $g_3(\cdot)$.
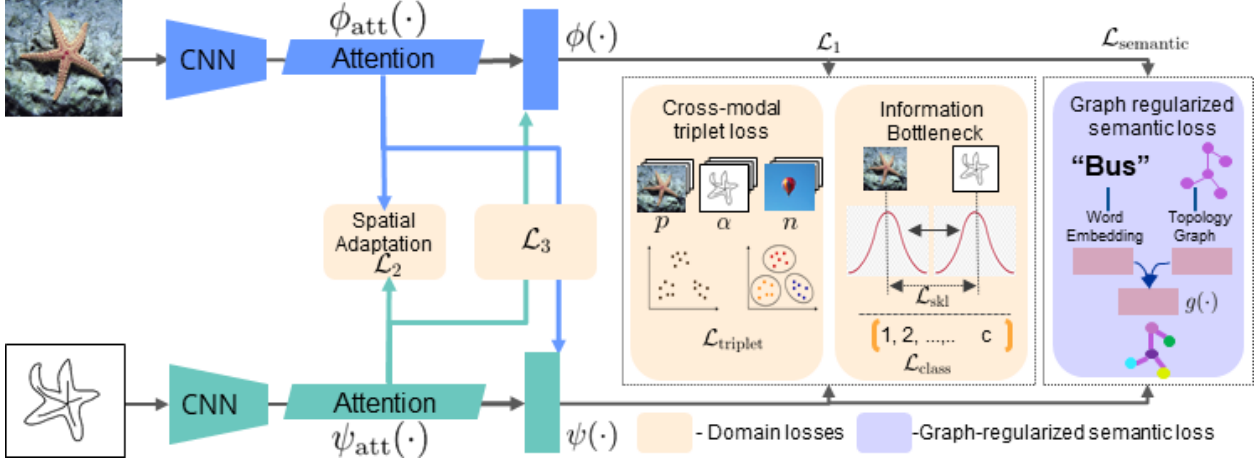
Figure 2: A depiction of our BDA-SketRet architecture. The images and sketches undergo a bi-level domain adaptation at the outputs of $(\phi_{\text{att}}, \psi_{\text{att}})$ and (ϕ, ψ), respectively. The important learning objectives: (i) domain losses, (ii) cross-modal multi-level reconstruction loss, and (iii) neighborhood preserving semantic projection loss are shown and how they are evaluated in the network. In the semantic graph, thicker edges means more discriminativeness. In the semantic graph, thicker edges means more discriminativeness. 0/1 represent the domain labels.

There are mainly three learning objectives that together train BDA-SketRet, namely, (i) spatial and semantic domain losses (ii) cross-modal multi-level reconstruction loss, and (iii) neighborhood preserving semantic projection loss. They are detailed in the following subsections.

4.1. Multi-level Sketch Image Alignment

The semantic adaptation works on the final features obtained from the last dense layer (ϕ and ψ) of the backbone, while the spatial adaptation is carried out on the mid-level feature maps obtained from the final convolutional layer (ϕ_{att} and ψ_{att}).

(i) Semantic Adaptation. We aim to maximise the information between the probability distributions $P_\alpha = \text{P}(\psi(\alpha)|\alpha)$ and $P_p = \text{P}(\phi(p)|p)$, and minimise the information they share with $P_n = \text{P}(\phi(n)|n)$. We employ a posterior distribution matching method to generate a latent space in which α and p are as aligned as possible, while α and n are sufficiently disjoint. We model the probability distributions P_α, P_p, and P_n by appropriate normal distributions.

We follow the deep variational information bottleneck framework (Alemi et al., 2017; Federici et al., 2020) where the classification task is considered for data from both the visual modalities with the combined cross-entropy loss as $\mathcal{L}_{class}$. The aim is to maximise information between latent variables and output space by minimising cross entropy loss between the distributions, and minimises information between source and latent variables by using the standard normal distribution as a variational approximation to their marginal. We restrain this latent space to be a standard normal distribution. We minimize the symmetric KL divergence between P_α and P_p is to reduce their relative entropy. Apart from reducing the distance between P_α and P_p, our framework also requires that P_α and P_n be disjoint in the feature space. For this, we define $\mathcal{L}_{skl}(P_\alpha, P_{p/n})$ as $\frac{1}{2}\big(D_{\text{KL}}(P_\alpha \| P_{p/n}) + D_{\text{KL}}(P_{p/n} \| P_\alpha)\big)$. Hence, we introduce a novel symmetric KL divergence triplet loss $\mathcal{L}_{t\text{-skl}}$ as,

$$\mathcal{L}_{t\text{-skl}} = \beta \max\{0, \mathcal{L}_{skl}(P_\alpha, P_p) - \mathcal{L}_{skl}(P_\alpha, P_n) + \lambda\} \quad (1)$$

Here β and λ are hyper-parameters which denote the weight of SKL divergence and the margin respectively. By definition of the loss function, our goal is to bring P_α and P_p closer while maintaining a difference of at least λ from the distance between P_α and P_n.

Along with $\mathcal{L}_{t\text{-skl}}$, we also use instance-based triplet loss $\mathcal{L}_{triplet}$ here so that the cross-modal data become class-wise dense clusters in the embedding space. They complement each other in obtaining a dense feature space highlighting only the domain-shared features while neglecting any artifact. While $\mathcal{L}_{t\text{-skl}}$ maximizes the information between latent variables and output space, $\mathcal{L}_{triplet}$ makes the cross-modalities dense in the embedding space. The cross-modal triplet loss aims to bring the same class sample $\phi(p)$ from the image modality closer to a given sketch anchor $\psi(\alpha)$ while pushing the negative image sample $\phi(n)$ far from $\psi(\alpha)$ at least by a margin of μ as per Eq. 2 using the Euclidean distance metric D.

$$\mathcal{L}_{triplet} = \min_{\phi,\psi} \mathbb{E}_{\alpha \in \mathcal{B}^s, p, n \in \mathcal{A}^s}[\max\{0, \mu + D(\psi(\alpha), \phi(p)) \\ - D(\psi(\alpha), \phi(n))\}] \quad (2)$$

Therefore, the combined semantic adaptation loss is: $\mathcal{L}_1 = \min_{\phi,\psi,h} \mathcal{L}_{t\text{-skl}} + \mathcal{L}_{triplet} + \mathcal{L}_{class}$.

(ii) Spatial Adaptation encourages the learning of abstract local spatial concepts common to $\mathcal{A}^s$ and $\mathcal{B}^s$. This is done by adversarially adapting the spatially average-pooled feature-map outputs of $\phi_{\text{att}}(p)$ and $\psi_{\text{att}}(\alpha)$. Most adversarial domain adaptation frameworks employ a binary domain classifier $l(\cdot)$. We observed that sketches and images are extremely disparate in terms of spatial and spectral properties. Therefore, it is fairly unchallenging for a discriminator to learn distributions corresponding to domains promptly, defeating the purpose of adversarial training. The root of this problem is the hard decision boundaries we set for the domain classifier in the min-max optimisation. We argue that the margin between the decision boundaries is still substantially large for the generator to overcome, given the inherently huge domain gap. Therefore, we

introduce a pseudo-decision boundary between spatial features of α and p that helps the generator to surpass the margin.

This decision boundary introduces a certain degree of domain invariance before projecting them onto the latent space, leading to stable training, i.e. $l(x) = 0.5$, if $x \in \mathcal{A}^s$ or $x \in \mathcal{B}^s$. In an adversarial setup, the generator competes with the discriminator to increase $P(x \in \mathcal{A}^s)$ to be greater than 0.5 and $P(x \in \mathcal{B}^s)$ to be less than 0.5. The decision boundary is symmetrical to both the domains in the feature space, rendering 0.5 as a good value for a pseudo decision boundary. So the local domain adaptation is given as:

$$\mathcal{L}_2 = \min_{\phi_{\text{att}}, \psi_{\text{att}}} \max_l \mathbb{E}_{\alpha \in \mathcal{B}^s, p \in \mathcal{A}^s} [0.5 \log(1 - l(\psi_{\text{att}}(\alpha)))$$
$$+ 0.5 \log l(\phi_{\text{att}}(p))] \quad (3)$$

(iii) Cross-modal multi-level reconstruction. Although $\mathcal{L}_2$ adapts the intermediate feature maps of $\mathcal{A}^s$ and $\mathcal{B}^s$, it does not guarantee that the feature maps of a given modality are aware of the final latent feature distributions for the other modality. In this regard, it is essential that the high-level feature distributions P_α and P_p characterised by domain-independence do not completely lose out class-related information about their conjugate modality i.e. α and p. To better equip the feature maps with the cross-modal information, we introduce the class-wise cross-modal reconstruction loss using generative modelling. Specifically, the cross-modal encoder-decoder modules $\mathcal{V}_\alpha = (\mathcal{V}_\alpha^e, \mathcal{V}_\alpha^d)$ and $\mathcal{V}_p = (\mathcal{V}_p^e, \mathcal{V}_p^d)$ reconstruct the latent feature embedding of sketch anchor $\psi(\alpha)$ given the outcome of $\phi_{\text{att}}(p)$ and vice-versa. Both $\mathcal{V}_\alpha^e$ and $\mathcal{V}_p^e$ are designed to be stochastic encoders and their outputs follow the standard normal distributions as per the principles of variational learning. Let D_{KL} be the Kullback-Leibler divergence, referred to as equation 4.

$$\mathcal{L}_{\text{KL}}(\mathcal{V}, F, x) = D_{\text{KL}}(q(\mathcal{V}(F(x))) \| \mathcal{N}(0, 1)) \quad (4)$$

The cross-modal reconstruction loss is given as follows.

$$\mathcal{L}_{\text{rec}}^1 = \left\| \mathcal{V}_p(\phi_{\text{att}}(p)) - \psi(\alpha) \right\|^2 + \mathcal{L}_{\text{KL}}(\mathcal{V}_p^e, \phi_{\text{att}}, p)$$
$$\mathcal{L}_{\text{rec}}^2 = \left\| \mathcal{V}_\alpha(\psi_{\text{att}}(\alpha)) - \phi(p) \right\|^2 + \mathcal{L}_{\text{KL}}(\mathcal{V}_\alpha^e, \psi_{\text{att}}, \alpha)$$
$$\mathcal{L}_3 = \min_{\mathcal{V}_p, \mathcal{V}_\alpha, \phi, \psi} \mathbb{E}_{\alpha \in \mathcal{B}^s, p \in \mathcal{A}^s} [\mathcal{L}_{\text{rec}}^1 + \mathcal{L}_{\text{rec}}^2] \quad (5)$$

The overall domain loss is given by equation 6.

$$\mathcal{L}_{\text{domain}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \quad (6)$$

4.2. Neighborhood Preserving Semantic Projection Loss

The semantic side information is obtained by taking various combinations of text-based and hierarchical word embeddings for the category names. For the distributed word-vector models, we consider the pre-trained Word2Vec (Mikolov et al., 2013) and fasttext (Bojanowski et al., 2017) while the Jiang-Conrath (Jiang and Conrath, 1997) and path similarity are used for the latter.

As aforementioned, the neighborhood information is severely affected when the semantic prototypes are directly projected onto the latent space using non-linear dense layers, limiting the overall performance. Hence, we desire to ensure the

latent space to mimic the neighborhood information of the original semantic space, in addition to being discriminative. We project the semantic prototypes $\mathcal{W}^s$ together with the topology information of the original semantic space to the shared latent space. The topology information, which is found to bring in a regularization effect into the latent space is encapsulated in the weighted semantic adjacency matrix $\Gamma_{|C^s| \times |C^s|}$ defined by the pairwise cosine dissimilarity among the semantic prototypes of the seen classes. Ideally, the outputs of $g_1(\mathcal{W}^s)$ and $g_2(\Gamma, \mathcal{W}^s)$ are concatenated and subsequently projected onto the latent space by another MLP $g_3(\cdot)$: $g(\mathcal{W}^s, \Gamma) = g_3([g_1(\mathcal{W}^s), g_2(\Gamma, \mathcal{W}^s)])$ where $[\cdot, \cdot]$ defines the vector concatenation operation. The semantic reconstruction loss brings $\phi(p)$ and $\psi(\alpha)$ closer to the projected class embedding $g(w^+, \Gamma)$ while maximizing the divergence between $\phi(n)$ and $g(w^+, \Gamma)$. This is accomplished through the graph regularized semantic loss $\mathcal{L}_{\text{semantic}}$ as follows,

$$\mathcal{L}_{\text{semantic}} = \min_{g, \phi, \psi} \mathbb{E}_{\alpha \in \mathcal{B}^s, p, n \in \mathcal{A}^s, w^+ \in \mathcal{W}^s} [S(\psi(\alpha), g(w^+, \Gamma), 1)$$
$$+ S(\phi(p), g(w^+, \Gamma), 1) + S(\phi(n), g(w^+, \Gamma), 0)]$$

where the distance S between the vectors $(\mathbf{x}, \mathbf{y})$ is defined in terms of the cosine distance for a given threshold t as, $S(\mathbf{x}, \mathbf{y}, t) = \frac{1}{2}(t - \frac{\mathbf{x}\mathbf{y}^T}{\|\mathbf{x}\|\|\mathbf{y}\|})$. The overall objective function for the BDA-SketRet framework can now be put forward as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{domain}} + \mathcal{L}_{\text{semantic}}$.

5. Theoretical proof for IB-based alignment for tighter generalization bound

In this section, we provide a theoretical proof to show that the proposed semantic adaptation and information bottleneck framework generate a latent space with two properties: the projections of α and p are coalesced, while those of α and n are disjoint. To prove this, we show that the distance between probability distributions of α and n is always greater than the distance between α and p using the theory of learning from different domains.

Adopting the notation from (Ben-David et al., 2009), we denote error given by a hypothesis $h \in \mathcal{H}$ on a source domain $\mathcal{S}$ and target domain $\mathcal{T}$ by ϵ_S and ϵ_T respectively, where $\mathcal{H}$ defines the hypothesis space. The error is defined as the probability according to the distribution D that a hypothesis h disagrees with a true labeling function f.

Let us first consider the objective of generating fused P_α and P_p. Without any loss of generality, we fix the source domain to be $D_\alpha = \{x \in A^s\}$ and target domain to $D_p = \{x \in B^s | x$ belongs to class $w^+\}$. Translating our notations to the theoretical bound obtained in (Ben-David et al., 2009), we provide the bound for the error on target domain as

$$\epsilon_p \leq d_{\mathcal{H}\Delta\mathcal{H}}(P_\alpha, P_p) + \epsilon_\alpha(h) + \min_{h' \in \mathcal{H}} \epsilon_p(h') + \epsilon_\alpha(h') \quad (7)$$

The constraints on the latent distributions to be the standard normal distribution as per the variational upper bound principle, results in the following inequality by using the triangular

Table 1: Comparing our BDA-SketRet with SOTA on ZS-SBIR (top) and GZS-SBIR (bottom) on both the splits of Sketchy-extended (S1 and S2), TU Berlin-extended and Quickdraw-extended datasets. All models use VGG-16 feature backbone. The '-' represents the evaluation metrices which were not mentioned in the respective papers.

Model	Sketchy-ext (S2)		Sketchy-ext (S1)		TU Berlin-ext		Quickdraw-ext	
	P@ 200	mAP @200	mAP	P@ 100	mAP	P@ 100	mAP	P@ 100
ZSIH (Shen et al., 2018)	-	-	-	-	22.0	29.1	13.1	18.8
CVAE (Yelamarthi et al., 2018)	19.6	-	33.3	22.5	00.5	-	-	-
ZS-SBIR (Yelamarthi et al., 2018)	-	-	19.6	28.4	00.5	00.1	00.6	00.1
SEM-PCYC (Dutta and Akata, 2019)	37.0	45.9	34.9	46.3	29.7	42.6	**17.7**	25.5
Doodle2search (Dey et al., 2019)	37.0	46.1	-	-	10.9	-	07.5	-
Style-guide (Dutta and Biswas, 2019)	40.0	35.8	37.6	48.4	25.4	35.5	-	-
Style-guide+AMDReg (Dutta et al., 2020)	-	-	41.0	51.2	29.1	37.6	-	-
SEM-PCYC+AMDReg (Dutta et al., 2020)	-	-	39.7	49.4	33.0	47.3	-	-
BDA-SketRet (Ours)	**45.8**	**55.6**	**43.7**	**51.4**	**37.4**	**50.4**	15.4	**44.0**
ZS-SBIR (Yelamarthi et al., 2018)	-	-	14.6	19.0	00.3	00.1	00.2	00.1
SEM-PCYC (Dutta and Akata, 2019)	-	-	30.7	36.4	19.2	29.8	14.0	22.1
SEM-PCYC+AMDReg (Dutta et al., 2020)	-	-	32.0	39.8	24.5	30.3	-	-
Style-guide (Dutta and Biswas, 2019)	-	-	33.0	38.1	14.9	22.6	-	-
GZS-BDA-SketRet (Ours)	**22.6**	**33.7**	**33.8**	**41.3**	**25.1**	**35.7**	**15.4**	**28.6**

inequality for the d-divergence.

$$d_{\mathcal{H}}(P_\alpha, P_p) \le d_{\mathcal{H}}(P_\alpha, \mathbb{N}(0, \mathbb{I})) + d_{\mathcal{H}}(P_p, \mathbb{N}(0, \mathbb{I})) \quad (8)$$

Similarly, by virtue of the information bottleneck, we can again use the triangle inequality for $d_{\mathcal{H}}(P_\alpha, P_p)$ as

$$d_{\mathcal{H}}(P_\alpha, P_n) \le d_{\mathcal{H}}(P_\alpha, \mathbb{N}(0, \mathbb{I})) + d_{\mathcal{H}}(P_n, \mathbb{N}(0, \mathbb{I})) \quad (9)$$

By Subtracting Eq. 8 from 9, we get

$$d_{\mathcal{H}}(P_\alpha, P_n) - d_{\mathcal{H}}(P_\alpha, P_n) \le d_{\mathcal{H}}(P_p, \mathbb{N}(0, \mathbb{I})) - \\ d_{\mathcal{H}}(P_n, \mathbb{N}(0, \mathbb{I})) \quad (10)$$

We would like to point out that the RHS consists of terms that are minimized by the variational bottleneck framework. Therefore from Eq. 5 , under optimal conditions on $\mathcal{L}_{\text{t-skl}}$, $d_{\mathcal{H}}(P_\alpha, P_p) - d_{\mathcal{H}}(P_\alpha, P_n) \le 0$. Therefore, we can conclude that the generated latent space probability distributions follow $d_{\mathcal{H}}(P_\alpha, P_n) \ge d_{\mathcal{H}}(P_\alpha, P_p)$.

6. Experiments

Datasets. We validate the efficacy of the BDA-SketRet by performing experiments on the benchmark Sketchy-extended (Sangkloy et al., 2016), TU Berlin-extended (TUB) (Eitz et al., 2012), and the newly introduced QuickDraw-extended (Dey et al., 2019) datasets. Sketchy consists of 125 categories of unpaired sketch and photo images. We use the two conventional train-test splits. In split 1 (S1) we randomly select 25 classes as the unseen test data, while in split 2 (S2) we use $|C^s| : |C^u| = 104 : 21$ as mentioned in (Yelamarthi et al., 2018) where the 21 unseen classes are carefully chosen not to be part of ImageNet (Deng et al., 2009). For the remaining datasets, we follow the same protocol as (Dutta and Akata, 2020).

Training and Evaluation Protocol. We select around 5000 triplets in each training iteration based on the aforementioned triplet-mining protocol. μ (Eq. 2) and λ (Eq. 1) for Sketchy: (0.1, 0.1), Tu-Berlin: (1, 1), and QuickDraw: (1, 1), while β is set to 0.0001 for all the datasets. These parameters were estimated using grid search with cross-validation (Yelamarthi et al., 2018). We use batch-normalization and leaky-ReLU non-linearity after each of the layers to ensure a stable training. $\mathcal{L}_{\text{total}}$ is optimized using the stochastic gradient descent (SGD) with momentum as the optimizer with a mini-batch size of 32. An initial learning rate of 0.0001 and a momentum of 0.9 are set. We find that $\mathcal{L}_{\text{total}}$ converges for all the datasets within 50 epochs. We report the performance of BDA-SketRet in terms of mAP@all (mean average precision), mAP@200, P@100, and P@200, respectively, where P stands for precision.

Implementation Details. The feature backbone networks $\phi(\cdot)$ and $\psi(\cdot)$ are the ImageNet pre-trained VGG-16 model (Simonyan and Zisserman, 2014). Two modality specific spatial attention learning modules consisting of convolution kernels with sigmoid non-linearity are applied on the outputs of the final convolution layer (conv-5) of ϕ and ψ, and the network upto the attention blocks each producing 512 feature maps of size 7×7. The attention blocks are followed by three new dense layers which project the attended feature maps onto the final latent space with dimensions $\mathbb{R}^{256}$. Besides, a spatial average pooling across the channels is applied on the outputs of $(\phi_{\text{att}}, \psi_{\text{att}})$ to obtain a single channel feature map of resolution 7×7. The encoder and decoder modules of $\mathcal{V}_\alpha$ and $\mathcal{V}_p$ are single dense layers of 128-d. The local and global domain classifiers $l(\cdot)$ and $f(\cdot)$, and the multi-class category classifier $h(\cdot)$ are also one dense layer each. For the semantic projection network $g(\cdot)$, g_1 and g_3 are three dense layers each. g_2 is a graph convolution layer, followed by the pooling and flattening layers. During training, VGG-16 layers prior to conv-5 are frozen while the proposed layers are updated.

6

7. Results

We choose the following state of the art methods, ZSIH (Shen et al., 2018), SEM-PCYC (Dutta and Akata, 2019), Doodle2search (Dey et al., 2019), Style-guide (Dutta and Biswas, 2019), AMD regularizer with SEM-PCYC and Style-guide (Dutta et al., 2020), respectively, for analyzing our performance. While SEM-PCYC, and Style-guide are based on adversarial training, ZS-SBIR utilizes variational encoder-decoder networks. AMD regularizer helps in tackling the data/class imbalance between the training and test sets. The performance of our full BDA-SketRet model on the Sketchy dataset with its two splits, the Tu-Berlin dataset and Quickdraw-extended datasets in comparison to the state of the art (SOTA) is reported in table 1. The mAP value for ZS-SBIR in Sketchy S2 is 43.5. Similar to BDA-SketRet, these techniques report their performances using the VGG-16 (Simonyan and Zisserman, 2014) based feature backbone networks. We have not compared our results with (Xu et al., 2020b; Chaudhuri et al., 2020) as the subset of test data considered is different from ours. For Sketchy, split 2 is considered to be more difficult than split 1 as it consists of test classes which are unseen to the ImageNet pre-trained networks.

Amongst the other competing techniques, we find the inclusion of AMDreg boosts the performance of the baseline ZS-SBIR systems. In spite of this, BDA-SketRet beats SEM-PCYC + AMD by a considerable margin of $3-5\%$. The TU-Berlin dataset is challenging mainly due to the presence of class-wise as well as domain-wise data imbalance. The performance of other competing techniques are extremely low. In contrast, we achieve a boost of 5% over the existing literature. The Quick-Draw dataset is excessively large consisting of highly ambiguous sketches and is by far the most challenging dataset for ZS-SBIR. BDA-SketRet beats the P@100 value over the SOTA with a margin of 10%, while falls marginally to SEM-PCYC in the mAP value.

Similar to ZS-SBIR, BDA-SketRet showcase overall improved performance measures for GZS-SBIR for all the datasets. In particular, BDA-SketRet produces high P@100 values of 41.3 for Sketchy (S1), 35.7 for TU-Berlin, and 28.6 for QuickDraw which are at least 2.5 % more than the literary works. No prior approach report the GZS-SBIR score for S2 for Sketchy yet. However, we find that our performance in this case is substantially high.

We note that the existing techniques only report P@100 or P@200 but not both. However, we feel that an effective ZS-SBIR system should produce high scores for both the metrics together. In this section, we report the performance of all the datasets, including the two splits of the Sketchy dataset with all the evaluation metrics in table 2. The top part report the ZS:SBIR results, while the bottom part reports the GZS:SBIR.

7.1. Evaluating Effect Of Input Modalities.

In the ZS-SBIR setup, the significance of the semantic information is imperative in maneuvering the alignment of the

Table 2: Performance of the proposed BDA-SketRet with ZS-SBIR (top) and GZS-SBIR (bottom) on both the splits of Sketchy, TUB and Quickdraw datasets on all evaluation metrices.

Dataset	mAP	P@ 100	P@ 200	mAP @200
Sketchy (S 1)	43.7	51.4	45.7	56.9
Sketchy (S 2)	43.5	51.2	45.8	55.6
TU-Berlin	37.4	50.4	43.8	54.4
QuickDraw	15.4	44.0	35.5	34.6
Sketchy (S 1)	33.8	41.3	31.4	38.6
Sketchy (S 2)	22.7	25.1	22.6	33.7
TU-Berlin	25.1	35.7	32.9	33.3
QuickDraw	15.4	28.6	29.5	27.4

Table 3: Effects of different semantic information on the mAP value for TU-Berlin and Sketchy (split 2) datasets.

W2v	Fasttext	Path	Jin-Con	Sketchy	TUB
✓				41.1	37.4
	✓			**43.5**	36.5
		✓		40.2	32.9
			✓	41.9	**37.4**
✓		✓		40.1	31.4
✓			✓	40.7	37.1
	✓	✓		39.1	32.8
	✓		✓	39.7	34.6

multi-modal data in the latent space. Different models yield different topological alignment of the classes in the latent space, which effectively causes the similar classes to cluster in a short range, while pushing apart the faraway classes. We consider the individual textual (300-d) and hierarchical embeddings as well as their concatenations and report the mAP values in table 3 for both Sketchy and TU-Berlin. We observe that there is a variation of up to $4-6\%$ in the performance of BDA-SketRet by using different semantic information. It is found that the individual semantic spaces provide superior performance than their pairwise combinations as neighborhood topology may not be consistent in different semantic spaces. We obtain the best performance of Sketchy is using the fasttext model, while it is Jiang-Conrath for TU-Berlin produces the best performance with a mAP of 37.4.

7.2. Evaluating Effect Of Feature Backbones.

Further, different backbone networks have been utilized by a few existing techniques for ZS-SBIR. It is unjust to directly compare them with the rest of the literary works which exploit

Table 4: Comparison of different visual backbones on BDA-SketRet and the corresponding SOTA. * denotes mAP@200. SAKE uses a differently trained feature backbone.

	BDA-SketRet		State-of-the-Art		
Pretrain	Sketchy	TUB	Reference	Sketchy	TUB
VGG-16	43.5	37.4	—- Table 1 —-		
ResNet-50	40.1	33.2	SkechGCN (Zhang et al., 2020)	38.2	32.4
ResNet-152	43.0*	25.8*	SAN (Pandey et al., 2020)	24.0*	14.0*
SE-ResNet50	51.2*	41.0	SAKE (Liu et al., 2019)	49.7*	47.5

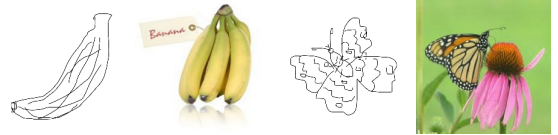Table 5: Ablation of loss functions and model components in terms of mAP for Sketchy (S2) and TU-Berlin. F denotes the full model. ($\mathcal{L}_{tri} = \mathcal{L}_{triplet}$, $\mathcal{L}_{sem} = \mathcal{L}_{semantic}$).

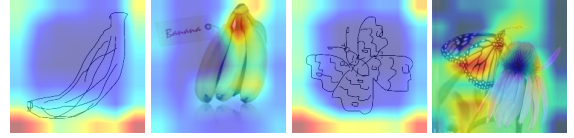|	Experimental set up	Sketchy	TUB
Losses	$\mathcal{L}_{sem} + \mathcal{L}_{tri}$	27.6	20.1
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_{t\text{-skl}}$	30.3	23.6
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_{t\text{-skl}} + \mathcal{L}_{class}$	33.4	27.7
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_{t\text{-skl}} + \mathcal{L}_{class} + \mathcal{L}_2$	36.9	33.5
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_{t\text{-skl}} + \mathcal{L}_{class} + \mathcal{L}_3$	40.1	34.8
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_2$	31.5	29.2
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_3$	33.2	27.0
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_2 + \mathcal{L}_3$	36.5	32.9
	$\mathcal{L}_{sem} + \mathcal{L}_{tri} + \mathcal{L}_{t\text{-skl}} + \mathcal{L}_{class} + \mathcal{L}_2 + \mathcal{L}_3$	**43.5**	**37.4**
Model	F w/o GCN	41.2	35.0
	F w/o attention block in local DA	38.6	32.5
	F w/o local attention & GCN	34.2	32.8

the conventional VGG-16 framework. Hence, we deploy different encoder networks to train BDA-SketRet and compare with the respective approaches (table 4) to provide a base-lining for the future endeavors. Similar to the semantic information, the chosen visual feature encoder affects the model performance considerably. SkechGCN (Zhang et al., 2020) considers the ResNet-50 (He et al., 2016) architecture while SAN (Pandey et al., 2020) utilizes the ResNet 152 (He et al., 2016), both pre-trained on the Imagenet dataset. We also test the performance of our framework using the trending SE-ResNet-50 feature extractor. SAKE (Liu et al., 2019) uses a conditional-SE-ResNet 50 (Hu et al., 2018) architecture, while using an auxiliary task to approximately map each image in the training set to the ImageNet semantic space. SE-ResNet is different from CSE-ResNet as it does not use any conditional variable. Similarly, (Thong et al., 2020) follow a different evaluation protocol from the remaining literature. While the other works use the entire seen classes along with the unseen classes for the GZS-SBIR experiments, in this paper the authors claim that they just use 20% of the samples from the seen classes for evaluation. Hence, a direct comparison of results with (Thong et al., 2020; Liu et al., 2019) may not be fair. Apart from these two, overall it can be observed that BDA-SketRet beats the concerned techniques consistently when adopting the respective visual feature extractors.

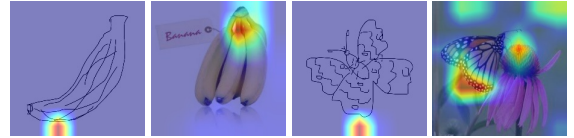### 7.3. Ablation Studies And Qualitative Results

Ablation of model components: The full model consists of a group of sub-modules, each contributing in its own way to enhance the performance. In the ablation analysis, the baseline network comprises of $\mathcal{L}_{triplet} + \mathcal{L}_{semantic}$. The global adaptation is performed on the latent features to reduce the domain-gap between the data from the two modalities by increasing the domain confusion. This improves the performance marginally, as seen from table 5 . This is expected to yield a class-wise overlapping embedding space for sketches and images. Simply adding the binary domain classifier without the label classifier leads to mode collapse. To avoid this and ensure class-wise discriminativeness, we add the full $\mathcal{L}_1$ loss and observe an increase in the overall performance. We then append the network with the local adaptation module applied on the intermediate
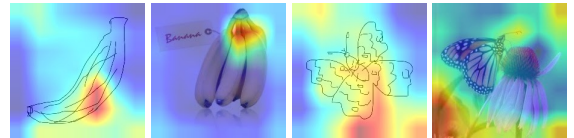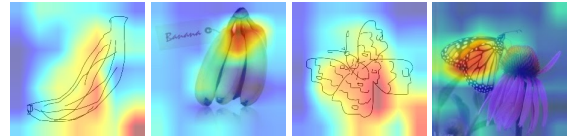

(a) Sample sketch and photo images.

(b) Grad-CAM plots highlighting the ROI on the model without $\mathcal{L}_{dom}^{global}$, $\mathcal{L}_{t\text{-skl}}$, $\mathcal{L}_2$ and $\mathcal{L}_3$.

(c) Grad-CAM plots highlighting the ROI on the model without $\mathcal{L}_{t\text{-skl}}$, $\mathcal{L}_2$ and $\mathcal{L}_3$ losses.

(d) Grad-CAM plots highlighting the ROI on the model without $\mathcal{L}_2$ and $\mathcal{L}_3$ losses.

(e) Grad-CAM plots highlighting the ROI when trained with the full model.

Figure 3: Ablation of domain adaptation stages using Grad-CAM plots to highlight the region of interest (ROI).

feature maps to highlight important local constructs common to both the modalities. When we add the cross-modal reconstruction modules, we observe significant improvements in the results (43.5 / 37.4). To study the individual contributions of $\mathcal{L}_2$ and $\mathcal{L}_3$, we use them individually and in conjunction with each other to the baseline model. A clear fall of about $7 - 10\%$ is observed, highlighting the contribution of the semantic adaptation module. As evident from table 5, the full model incurs a boost of $12 - 13\%$ on the mAP values for both the datasets than the baseline, ranging from 27.6 to 43.5 in the Sketchy and 20.1 to 37.4 in the TU-Berlin.

Further, we study the effects of the GCN module in $g(\cdot)$ and the spatial attention layers in ϕ_{att} and ψ_{att}, respectively. We observe a marginal performance drop of $1 - 2\%$ when the GCN layer is removed from the full BDA-SketRet. Similarly, the attention module is crucial in highlighting the domain-invariant mid-level features and BDA-SketRet without the attention layers is found to marginally degrade the performance. We also look into the effect of removing the GCN module for the GZS-SBIR experiments and notice a drop of 10% in mAP.
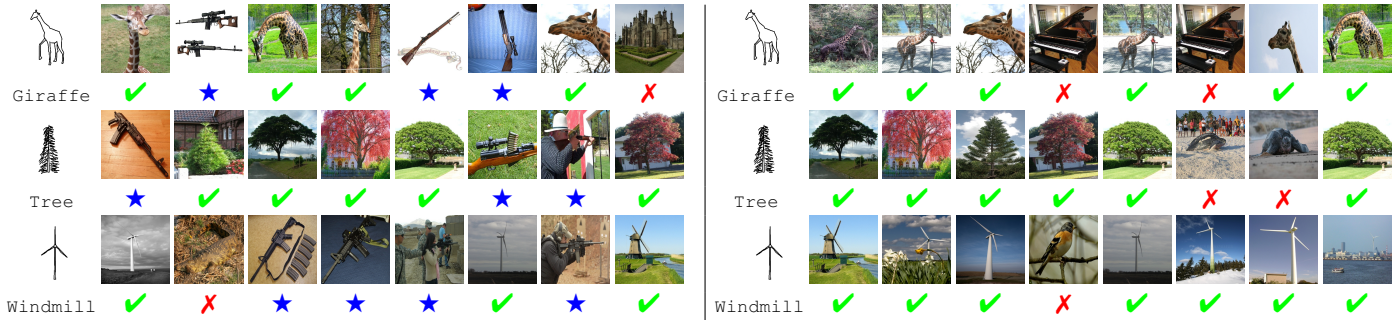
Figure 4: Top-8 retrieval instances for a few sketch queries from the Sketchy dataset without using the local adaptation on the left and BDA-SketRet on the right. The green checks denote correctly retrieved classes, while the red crosses denote incorrect class images. The blue stars denote the hub instances.

Qualitative analysis of hubness. Hubness problem (Radovanovic et al., 2010) occurs when a model has a training bias and retrieves images only from a subset of the available categories. This occurs as some embedding vectors of images (also called as "hubs") appear in the nearest neighborhood of many test query sketches. An ineffective feature alignment between the visual modalities may trigger the generation of hubs. In order to show the role of domain losses in retrieving more discriminative image samples, we first train the model with just the adversarial global adaptation (without $\mathcal{L}_2$ and $\mathcal{L}_3$), followed by training the entire BDA-SketRet model. In the first case, we notice the presence of hubs. Precisely, the left column of Fig. 4 shows a scenario where the instances of `rifle` class are retrieved for nearly many query samples as its embeddings are cluttered in the feature space with multiple classes. This adversely affects the overall performance. In the second case, no particular class is visibly found to clutter the retrieval results in the latent space. It depicts that by jointly using the global and local adaptations and using cross-modal reconstruction modules, we achieve a hub-free retrieval results for the same set of queries.

Preservation of semantic space using GCN In this section, we aim to show that the leveraging a graph convolution network (GCN), we preserve the semantic topology in the embedding feature space. The semantic information has an important role to play for attaining an improved ZS-SBIR inference. Generally, the visual modalities are aligned to a fixed semantic space. While (Dutta and Akata, 2019) uses an encoded semantic space obtained via a semantic auto-encoder, methods of (Dey et al., 2019; Dutta and Biswas, 2019) reconstruct the original semantic vectors from the visual information. These methods could lead to a drastic change of the semantic topology in the embedding feature space. Fig. 5 (a) from the main manuscript shows the co-variance matrix of the 250 classes of the TU-Berlin dataset, while Fig. 5 (b) shows the co-variance matrix of the same after passing through a semantic auto-encoder. It can be seen that the semantic topology is visibly distorted, which would lead to improper mapping of visual and semantic space for the unseen classes.

In the proposed BDA-SketRet framework, we utilize graph CNN (Kipf and Welling, 2016) to encode the structural similarity among the different classes. We model an independent latent

(a) Original word embeddings. (b) Auto-encoded word embeddings in latent space. (c) GCN preserved word embeddings in latent space.
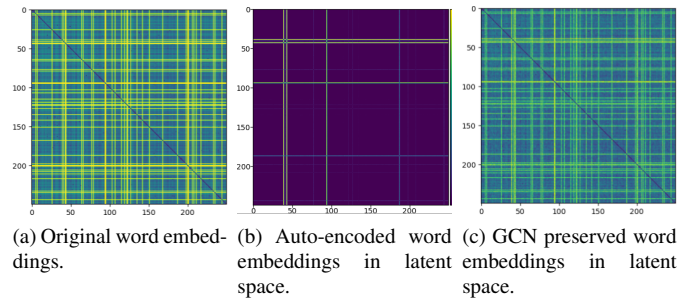
Figure 5: Co-variance matrices of the word embeddings under different experimental set-ups.

space given the original modalities but constraint the space to be influenced by the characteristics of the original semantic space. Fig. 5 (c) shows the co-variance matrix of the 250 classes of the TU-Berlin dataset. It can be seen that the semantic topology is visibly preserved in the embedding space. In effect, this helps us in obtaining a better mapping of the visual and semantic space for the unseen classes.

7.4. Grad-CAM Visualization

Gradient-weighted class activation mapping (Selvaraju et al., 2017) (Grad-CAM) primarily uses the gradients of the target class at the final convolution layer to synthesize an intermediate localization map which highlights the most important regions in the image. It effectively helps in displaying the region which gets the most importance for any particular target-class.

Effect of bi-level adaptation. We ablate among the various domain adaptation modules in the network and provide a visualization of the same in Fig. 3 and show the Grad-CAM plots of both the models for a few sketch and photo images.

In Fig. 3 (b), we train the model without the $\mathcal{L}_{\text{dom}}^{\text{global}}$, $\mathcal{L}_{\text{t-skl}}$, $\mathcal{L}_2$ and $\mathcal{L}_3$ losses. Notice that this causes improper distribution of weights especially in the sketch images. The network primarily learns from the background of the sketch images, while for the photo images, the region of importance is scattered all over the image. Adding the $\mathcal{L}_{\text{dom}}^{\text{global}}$, relatively constricts the scattered region of importance to a more localized form. When we further go ahead and add the $\mathcal{L}_{\text{t-skl}}$ loss to the model, from Fig. 3 (d) we start noticing that the alignment of important region in the

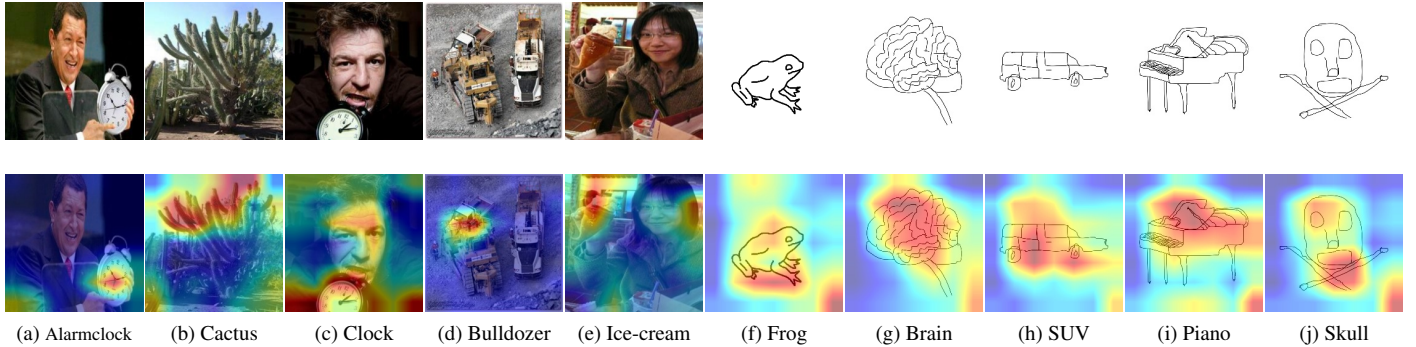|									
(a) Alarmclock	(b) Cactus	(c) Clock	(d) Bulldozer	(e) Ice-cream	(f) Frog	(g) Brain	(h) SUV	(i) Piano	(j) Skull

Figure 6: Grad-CAM plots of full model of a few sample photo and sketch images.

sketch images start improving. We can see that the full model produces better highlight to the local constructs.

This establishes our claim that the notion of fine-grained domain adaptation helps in obtaining a more discriminative latent space to combat hubness and negative knowledge transfer judiciously. The Grad-CAM plots displays the region which gets the most importance (ROI) when we train the model. We can see that the full model produces better highlight to the local constructs. Notice in the butterfly image, without the bi-level adaptation, the attention is splattered across the image and this leads to incorrect class assignment during retrieval. Using the full model of BDA-SketRet with bi-level adaptation, although the foreground consists of both the butterfly and the flower, the importance is layed properly on the concerned `butterfly` class.

Qualitative results. In Fig. 6 we show the GradCAM plots of a few sample photo and sketch images. From the photo images it can be seen that query class is properly highlighted in each image, keeping the remaining image as a background. For example, it can be seen from the `Alarmclock` images that the attended region is carefully just the clock, leaving out the human as the background. Similarly, the dendritic-structure of the `cactus` receives the most attention as they contribute more towards the overall recognition of its class. It can also be seen that out of the two power vehicles, the network correctly chooses the `bulldozer` and puts more weight on it.

Similarly, for the sketch classes it is seen that the highlighted region is the most important characteristic part of the sketch. For example, for `frog`, the webbed feet are the most important region, while for a `skull`, the cross-bones commonly drawn under skulls along with the hollow eye sockets conveyed the most information. It can be noted that since CNNs have a tendency to learn from image textures (Geirhos et al., 2018), the network provides more weights towards the image boundaries (as seen from Fig. 3 (b)). This is exactly the reason why sketch-based learning tasks are challenging. The proposed BDA-SketRet helps the network to learn the most important regions and their corresponding weights properly even from sketch images as shown in Fig. 6. In Fig. 7, we show the ZS-SBIR results for a few sample sketches from the Sketchy-extended dataset.

Table 6: Sensitivity analysis of hyper-parameters in terms of mAP value for TU-Berlin and Sketchy (split 2) datasets.

μ	λ	β	**Sketchy**	**TUB**
1	1	0.0001	41.8	**37.5**
0.1	1	0.0001	42.2	37.1
0.01	1	0.0001	42.7	36.7
0.1	0.1	0.0001	**43.5**	35.2
0.1	0.01	0.0001	40.6	32.0
1	1	0.001	42.5	35.9
1	1	0.01	38.5	33.4
1	1	0.00001	41.6	36.8

7.5. Sensitivity To Hyper-parameters

In the proposed work, we have primarily 3 hyper-parameters, i.e., λ, β, and μ from the equations 1 and 2 of the main manuscript. For tuning the hyper-parameters, we follow the standard protocol of the zero-shot learning community by splitting the training data for cross-validation into training and validation data (psuedo-seen and psudeo-unseen split). The overall performance of the proposed framework by varying these parameters with some of the settings are shown in Table 6 for both Sketchy and TU-Berlin in terms of mAP values. We experiment the model performance by choosing different values in the range of 0.01 to 1 for μ and λ, while 0.00001 to 0.01 for β, as the $\mathcal{L}_{\text{t-skl}}$ loss is found to converge the fastest amongst the rest. We choose the combination of μ, λ, and β values that result in the best model performance.

8. Conclusions

We introduce a novel ZS-SBIR framework called BDA-SketRet in this paper. The main premise of our model is to perform improved alignment between the image and sketch features based on both the mid-level and high-level CNN based feature embeddings. Together, we introduce two generative cross-modal reconstruction modules to ensure the learning of robust modality-independent features. We further propose to project the semantic information into the shared latent space through a two-stream fusion network by jointly exploiting both the prototypes and the semantic class neighborhood. Overall, BDA-SketRet learns a discriminative and compact latent space and wisely tackles both the negative transfer and the hubness
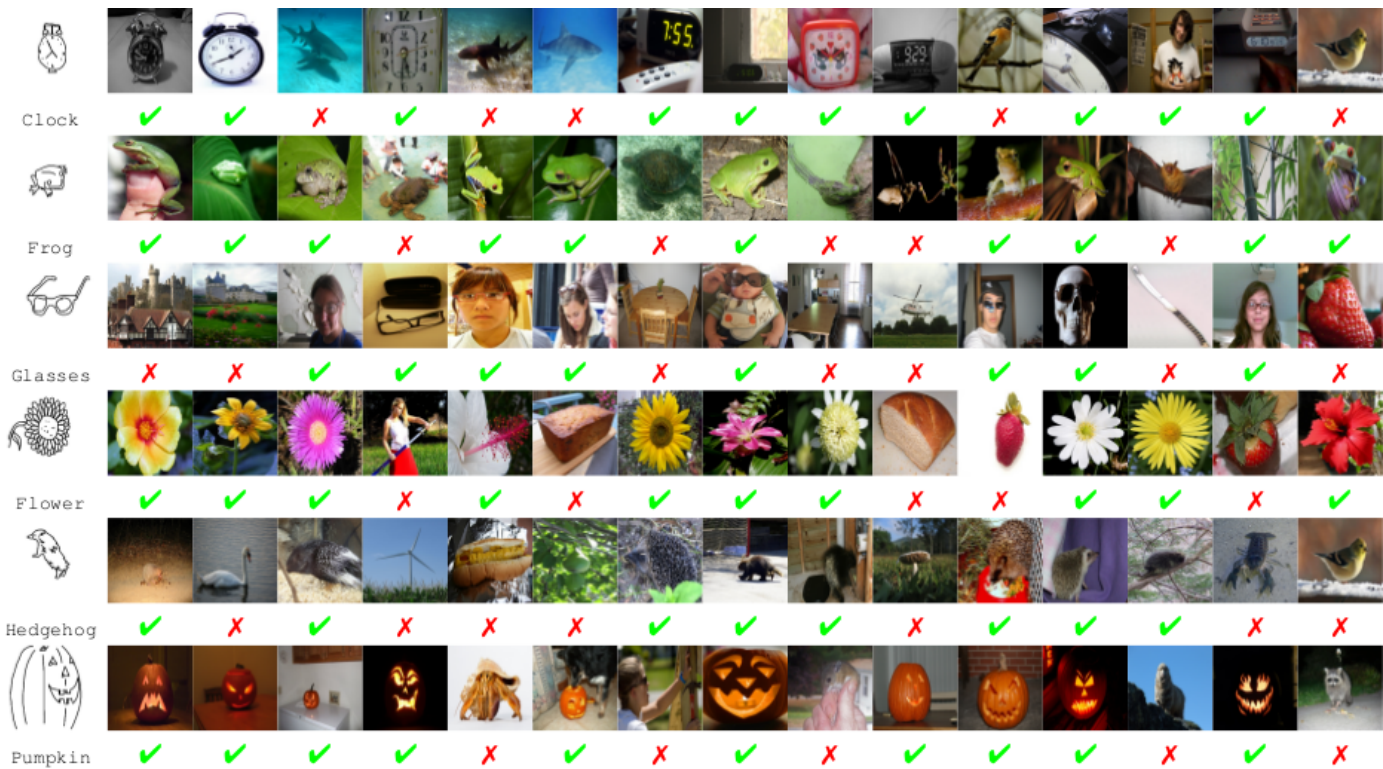
Figure 7: Top-15 retrieval instances for a few sketch queries from the Sketchy dataset using the full model. The green checks denote correctly retrieved classes, while the red crosses denote images from incorrect class. Notice that there are no hub instances generated here.

issues of domain adaptation and ZSL, respectively. Experimentally, we outperform the recent techniques in all the performance metrics on all the existing datasets. We are currently interested in extending BDA-SketRet to support the paradigm of lifelong learning.

References

Alemi, A., Fischer, I., Dillon, J., Murphy, K., 2017. Deep variational information bottleneck, in: ICLR.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features, in: ECCV.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F.C., Vaughan, J.W., 2009. A theory of learning from different domains. Machine Learning 79, 151–175.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. ACL .

Bui, T., Ribeiro, L., Ponti, M., Collomosse, J., 2017. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. CVIU .

Chaudhuri, U., Banerjee, B., Bhattacharya, A., Datcu, M., 2020. Crossatnet - a novel cross-attention based framework for sketch-based image retrieval. IVC .

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: CVPR.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: CVPR.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: CVPR.

Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.Z., 2019. Doodle to search: Practical zero-shot sketch-based image retrieval, in: CVPR.

Dutta, A., Akata, Z., 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval, in: CVPR.

Dutta, A., Akata, Z., 2020. Semantically tied paired cycle consistency for any-shot sketch-based image retrieval. IJCV .

Dutta, T., Biswas, S., 2019. Style-guided zero-shot sketch-based image retrieval., in: BMVC.

Dutta, T., Singh, A., Biswas, S., 2020. Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir, in: ECCV.

Dutta, T., Singh, A., Biswas, S., 2020. Styleguide: Zero-shot sketch-based image retrieval using style-guided image generation. IEEE TM .

Eitz, M., Hays, J., Alexa, M., 2012. How do humans sketch objects? ACM TOG .

Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M., 2010. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE TVCG .

Federici, M., Dutta, A., Forré, P., Kushman, N., Akata, Z., 2020. Learning robust representations via multi-view information bottleneck, in: ICLR.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2018. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 .

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: CVPR.

Hu, R., Collomosse, J., 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. CVIU .

Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv .

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., Xing, E.P., 2019. Rethinking knowledge graph propagation for zero-shot learning, in: CVPR.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks, in: ICLR.

Lei, J., Song, Y., Peng, B., Ma, Z., Shao, L., Song, Y.Z., 2019. Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology 30, 3226–3237.

Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L., 2017. Deep sketch hashing: Fast

free-hand sketch-based image retrieval, in: CVPR.

Liu, Q., Xie, L., Wang, H., Yuille, A.L., 2019. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval, in: ICCV.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: Proceedings of the seventh IEEE international conference on computer vision.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv .

Narasimhan, M., Lazebnik, S., Schwing, A., 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering, in: NeurIPS.

Pandey, A., Mishra, A., Verma, V.K., Mittal, A., Murthy, H., 2020. Stacked adversarial network for zero-shot sketch based image retrieval, in: WACV.

Qi, Y., Song, Y.Z., Zhang, H., Liu, J., 2016. Sketch-based image retrieval via siamese convolutional neural network, in: ICIP.

Radovanovic, M., Nanopoulos, A., Ivanovic, M., 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. JMLR .

Romera-Paredes, B., Torr, P., 2015. An embarrassingly simple approach to zero-shot learning, in: ICML.

Saavedra, J.M., 2014. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo), in: ICIP.

Sangkloy, P., Burnell, N., Ham, C., Hays, J., 2016. The sketchy database: learning to retrieve badly drawn bunnies. ACM TOG .

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: ICCV.

Shen, Y., Liu, L., Shen, F., Shao, L., 2018. Zero-shot sketch-image hashing, in: CVPR.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv .

Song, J., Yu, Q., Song, Y.Z., Xiang, T., Hospedales, T.M., 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval, in: ICCV.

Thong, W., Mettes, P., Snoek, C.G., 2020. Open cross-domain visual search. CVIU .

Tishby, N., Zaslavsky, N., 2015. Deep Learning and the Information Bottleneck Principle, in: ITW.

Wang, M., Wang, C., Yu, J.X., Zhang, J., 2015. Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. VLDBE .

Xian, Y., Lorenz, T., Schiele, B., Akata, Z., 2018. Feature generating networks for zero-shot learning, in: CVPR.

Xian, Y., Schiele, B., Akata, Z., 2017. Zero-shot learning-the good, the bad and the ugly, in: CVPR.

Xu, F., Yang, W., Jiang, T., Lin, S., Luo, H., Xia, G.S., 2020a. Mental retrieval of remote sensing images via adversarial sketch-image feature learning. IEEE Transactions on Geoscience and Remote Sensing 58, 7801–7814. doi:10.1109/TGRS.2020.2984316.

Xu, X., Yang, M., Yang, Y., Wang, H., 2020b. Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval, in: IJCAI.

Yang, X., Tang, K., Zhang, H., Cai, J., 2019. Auto-encoding scene graphs for image captioning, in: CVPR.

Yelamarthi, S.K., Reddy, S.K., Mishra, A., Mittal, A., 2018. A zero-shot framework for sketch based image retrieval, in: ECCV.

Yu, Q., Yang, Y., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., 2017. Sketch-a-net: A deep neural network that beats humans. IJCV .

Zhang, J., Shen, F., Liu, L., Zhu, F., Yu, M., Shao, L., Shen, H.T., Van Gool, L., 2018. Generative domain-migration hashing for sketch-to-image retrieval, in: ECCV.

Zhang, Z., Zhang, Y., Feng, R., Zhang, T., Fan, W., 2020. Zero-shot sketch-based image retrieval via graph convolution network., in: AAAI.