This CVPR 2020 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# A Simplified Framework for Zero-shot Cross-Modal Sketch Data Retrieval

Ushasi Chaudhuri
Indian Institute of Technology
Bombay, India
ushasi@iitb.ac.in

Biplab Banerjee
Indian Institute of Technology
Bombay, India
getbiplab@gmail.com

Avik Bhattacharya
Indian Institute of Technology
Bombay, India
avikb@csre.iitb.ac.in

Mihai Datcu
German Aerospace Center (DLR)
Germany
mihai.datcu@dlr.de

## Abstract

*We deal with the problem of zero-shot cross-modal image retrieval involving color and sketch images through a novel deep representation learning technique. The problem of a sketch to image retrieval and vice-versa is of practical importance, and a trained model in this respect is expected to generalize beyond the training classes, e.g., the zero-shot learning scenario. Nonetheless, considering the drastic distributions-gap between both the modalities, a feature alignment is necessary to learn a shared feature space where retrieval can efficiently be carried out. Additionally, it should also be guaranteed that the shared space is semantically meaningful to aid in the zero-shot retrieval task. The very few existing techniques for zero-shot sketch-RGB image retrieval extend the deep generative models for learning the embedding space; however, training a typical GAN like model for multi-modal image data may be non-trivial at times. To this end, we propose a multi-stream encoder-decoder model that simultaneously ensures improved mapping between the RGB and sketch image spaces and high discrimination in the shared semantics-driven encoded feature space. Further, it is guaranteed that the class topology of the original semantic space is preserved in the encoded feature space, which subsequently reduces the model bias towards the training classes. Experimental results obtained on the benchmark Sketchy and TU-Berlin datasets establish the efficacy of our model as we outperform the existing state-of-the-art techniques by a considerable margin.*

## 1. Introduction

Research has been at its best in the last few years due to the advancement in high-computing technologies and the accumulation of very large scale data. This is also partly due to improved technologies for cost-effective and vast multitudes of varieties of sensors. Due to this, the same data can be described in different formats and representations. Since each of these representations often encompasses the features from the same object, inter-modal data retrieval between these representations finds a huge application. This is because a few forms of data representation could be more informative than the others, while some might be more visualizable for better understanding.

One such minimalistic representation of visual data is in the form of sketches. With the availability of cheap styluses, drawing quick sketches can be made in the nick of time. Sketch-based image retrieval (SBIR) finds numerous applications nowadays in forensic studies for police investigations made using criminal sketch queries. In SBIR, we query a sketch and retrieve corresponding nearest $k$ neighbors from a large-scale image database. This idea can be extended to make a cross-modal retrieval framework; wherein one can incorporate both SBIR and IBSR (image-based sketch retrieval). However, it can be noted that since sketches lack texture information, learning using CNNs become extremely difficult as CNNs have been seen to have an inherent bias towards learning from the image textures [9]. Hence, SBIR and IBSR are challenging problems that are highly researched upon in recent times.

Out of the few limited works that have been dome in cross-modal retrieval (CMR), only a few have worked on zero-shot learning (ZSL) based CMR. A few notable works in ZSL based SBIR have started coming up recently. The idea of ZSL is to train a network for a select number of training classes *seen classes* for which the training samples are available, along with its class attribute information. The aim is to train the network so well that while testing if we provide an entirely new and non-overlapping class or *unseen class* data samples along with its attribute information,

the network should be able to classify it into that class correctly. The difficulty level shoots up when we deploy such a framework for a CMR architecture.

Most of the works in SBIR, which are done nowadays, use generative models. This is done by creating pseudo samples of unseen classes using their semantic information by incorporating discriminator and generator functions. However, the main problem with this type of work is, it is often challenging to train a stable network. This can happen if the derived feature space is not particularly sufficiently discriminative enough. This is difficult in sketches as sketches are a minimalistic representation of data and can be easily confused withing various classes. Some of the prominent works using generative networks are [21, 11].

**Our contributions:** Taking into account the problems as mentioned earlier and the gap in the literature, we propose a very simple yet novel zero-shot based cross-modal retrieval of sketch images. While the existing literature only focuses on retrieving the SBIR part, we make our network robust for both SBIR and IBSR problems. Single way network is easier to train as we address only half of the complete problem statement. The moment we encode the IBSR part to the model, the performance of the SBIR is partially compromised. Hence, the network is designed in such a way to provide maximum SBIR and IBSR performances. Our main contributions in this work are:

1) We use a cross-modal retrieval framework by constructing a shared discriminative embedding space for the data instances of different modalities. 2) We use a cross-reconstruction network to bridge the domain gap between the different modalities. 3) A cross-triplet loss for increasing the intra-class distance in the feature space, while reducing the inter-modal distance. 4) We extend the network for a zero-shot based CMR. 5) We experiment with the performance of the proposed model on the standard benchmarked large-scale databases, i.e., TU-Berlin and Sketchy, and obtain superior performance than the current state-of-the-art models. It is a reasonably easy model and very quickly trained.

## 2. Related Works

In this section, we discuss the main works in literature that have relevance in our domain. The literature survey part mainly comprises of CMR and ZSL in SBIR.

Few of the notable early works in the field of retrieval was mainly done by [6, 8]. These were non-learning based retrieval frameworks from a single modality of data. Soon with the popularity in learning-based techniques, many more works started coming up using pre-trained networks, like Resnet, GoogleNet, VggNets, etc. [22, 4, 2]. With the introduction to learning-based techniques and the accumulation of unprecedented volumes of data from different sensors, researchers started focusing more on cross-modal retrieval techniques. Some of the notable CMR works in literature are mostly image and text pair CMR, which has been quite extensively researched upon [16]. CMR also finds essential applications in RGB image to depth estimation in computer vision, SLAM problems, and depth estimation from LiDAR point clouds [23]. Some works have also been done on image and speech pair retrieval [3]. Recently, with the advancement in touch and sensor technology, people have been researching on sketch-based image retrieval problems (SBIR). However, the present work done in literature does not involve a cross-modal retrieval, rather a single way retrieval framework only [27, 29, 25, 12].

ZSL framework uses the concept of making a model learn from a set of seen classes. The idea is to make the model learn so well that on deployment, it can recognize even unseen classes (which is a disjoint set of classes from the seen class set). A detailed review of the literature in zero-shot has been provided in [7]. A conspicuous early work in ZSL was by [1], after which researches on ZSL gained acceleration. For associating the class features from one set of classes to their corresponding class label, we use a class-descriptive attribute or semantic information. The semantic information can be deployed in many ways. Few researchers have used visual semantic mappings [10], while some have used word2vec embeddings of the class labels [5], on the other hand, some used a complete sentence description of the image captions [19]. Most of the works in ZSL, which are done nowadays, use generative models. This is done by creating pseudo samples of unseen classes using their semantic information, by incorporating discriminator and generator functions [5, 11]. In ZSL-SBIR, a few of the notable works include [11, 21, 5, 14]. These models have been experimented on the standard benchmarked TU-Berlin and the Sketchy dataset and have shown to have achieved the current state-of-the-art results.

**How we are different?** There exists only a handful of works in SBIR. To the best of our knowledge, none of them have been extended to a bi-directional retrieval framework. Out of the few existing bidirectional-retrieval frameworks, their performance on such minimalistic data representation like sketch data remains unexplored. We design a bi-directional retrieval framework, with zero-shot learning for image-sketch retrieval. While most of the works in ZSL:SBIR use generative models by creating pseudo samples of unseen classes using their semantic information [5, 11], we have used a simple encoder-decoder strategy for training. These kinds of generator and discriminator based models are usually not very stable and are hence considered challenging to be trained. Most of the models in literature use complicated high-level attributes derived from multiple models for designing the semantic vector for the unseen classes. We, on the other hand, use just the word2vec word encoding of the label class and yet outper-

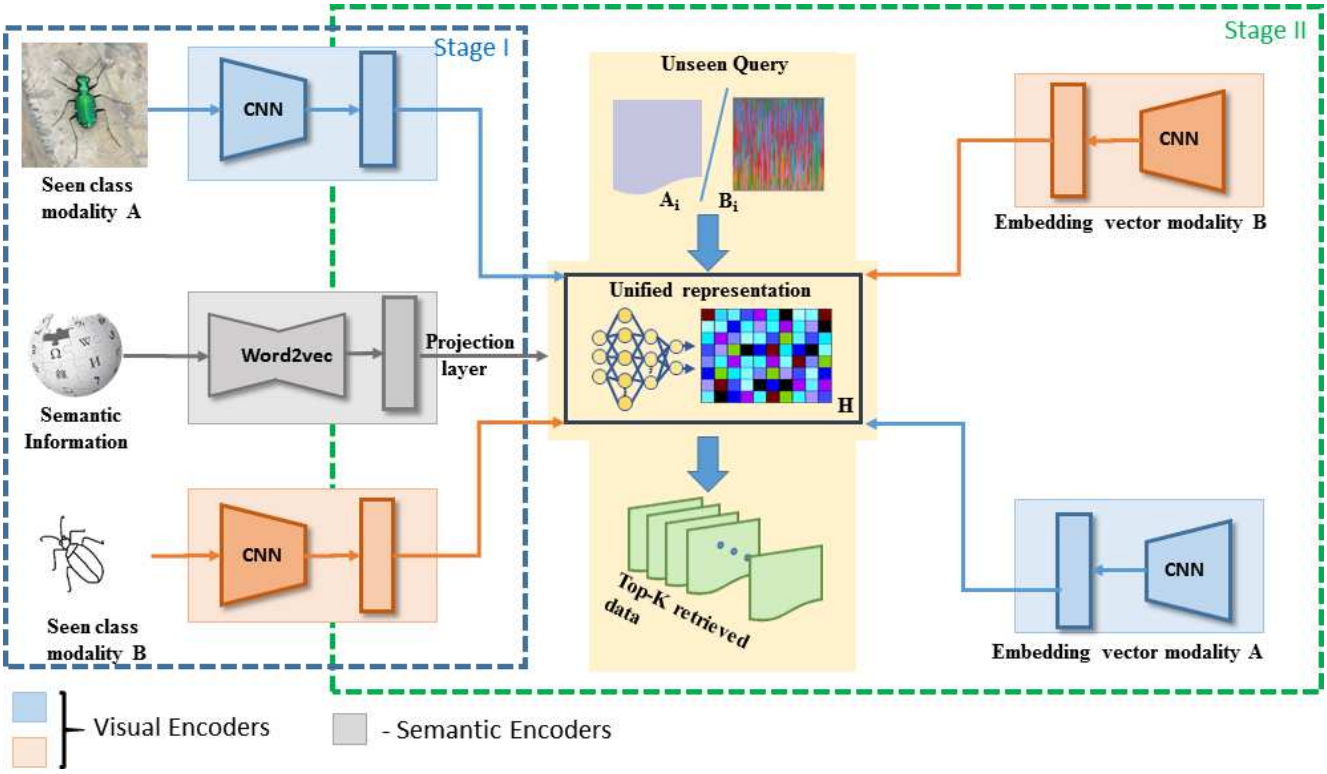


Figure 1. The Overall pipeline of the proposed ZSCMR architecture for a zero-shot retrieval from a cross-modal database, by projecting the data samples on a shared semantic space. The network is trained on the *seen* classes and tested on *unseen* classes. **H** represents the unified representation. With the help of the projection layer, retrieval from an unseen class is possible.

form all the current-state-of-the-art models for the same. We introduce a novel cross-triplet loss here, which helps primarily in boosting the CMR performance.

## 3. Methodology

**Preliminaries:** Let us denote two of the modalities corresponding to images and sketches as $\mathcal{A}$ and $\mathcal{B}$, respectively. The goal of the cross-modal retrieval setup is to retrieve similar images in $\mathcal{A}/\mathcal{B}$ given query images from the other modality $\mathcal{B}/\mathcal{A}$. For a ZSL framework, we partition the dataset into two subsets based on the classes, namely, the *seen* and the *unseen*. Both the sets are henceforth represented by $\mathcal{S} = \{\mathcal{A}^s, \mathcal{B}^s, \mathcal{Y}^s, \mathcal{Z}^s\}$ and $\mathcal{U} = \{\mathcal{A}^u, \mathcal{B}^u, \mathcal{Y}^u, \mathcal{Z}^u\}$, respectively, with the constraint that $(\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing)$. Note that $\mathcal{Y}^s$ defines the label-set for the seen classes and $Z^s$ represents the semantic class prototypes for the same (one prototype per class). Furthermore, a given instance for (un)seen set is denotes as $(a_i^{s/u}, b_i^{s/u}, y_i^{s/u}, z_i^{s/u})$. As aforementioned, the model is train only using $\mathcal{S}$ while $\mathcal{U}$ is only deployed during inference. Finally, our prime goal in zero-shot cross-modal retrieval is to design a mapping between the multi-modal visual space and the semantic space: ($\mathcal{A}^s$, $\mathcal{B}^s$) and $\mathcal{Z}^s$ such that given an $a_i^u \in \mathcal{A}^u$ and $\mathcal{Z}^u$, we can correctly retrieve similar images from $\mathcal{B}^u$ and vice-versa.

We detail the proposed methodology in this section. A depiction of the overall architecture of our model is illustrated in Figure 1.

### 3.1. Overall ZSCMR Architecture

To realize the ZSL framework on cross-modal data, we use a two-stage training process, described as follows:

1. We first model the two modality-specific classifiers given $\mathcal{A}^s$ and $\mathcal{B}^s$, respectively. We fine-tune an Imagenet pre-trained CNN model for both the modalities separately.

2. In the next stage, we introduce a multi-stream encoder-decoder based neural network model on carrying out the visual-semantic mapping task for ZSL based cross-modal retrieval.

**Stage 1:** Since our ZSL based retrieval model contains encoder branches corresponding to both the modalities, we prefer to pre-train these modality-specific encoder branches, which, in turn, offer better weight initialization while training the ZSL model. Typically, we fine-tune the pre-trained VGG and Resnet networks on $(\mathcal{A}^s, \mathcal{Y}^s)$ and

$(\mathcal{B}^s, \mathcal{Y}^s)$, respectively. The trained feature encoders in this way are directly utilized in the ZSL network separately.

**Stage 2:** In the second stage, we aim to learn a shared latent feature space **H** which ensures that the related data from $\mathcal{A}^s$ and $\mathcal{B}^s$ with identical labels should be closely aligned with the respective class prototype vectors from $\mathcal{Z}^s$ in the latent space. We propose a multi-stream encoder-decoder network for this purpose. In the following, we detail the design of this network. Broadly, the encoder part contains three branches: i) two of the branches corresponding to the visual modalities $\mathcal{A}$ and $\mathcal{B}$, respectively, while another branch which encodes the semantic information $\mathcal{Z}$.

**a) Visual Encoders:** For the visual data streams, we use encoders $f_A(;,\theta_A)$ and $f_B(;,\theta_B)$ to obtain the input features corresponding to $\mathcal{A}$ and $\mathcal{B}$ respectively. $\theta_A$ and $\theta_B$ denote the learnable parameters for both the encoder. Note that both $\theta_A$ and $\theta_B$ are initialized in stage 1. Both the output of the encoders yield feature embeddings with identical dimensionality so that they can be compared to subsequently define the shared latent space.

**b) Semantic Encoders:**

Likewise, we also learn the semantic encoder $f_Z(;,\theta_Z)$ with the learnable parameter set $\theta_Z$ for embedding the semantic information in $\mathcal{Z}$. Initially, a given $z \in \mathcal{Z}$ is represented in terms of the distributed word-vector embedding vectors corresponding to the semantic class names. Such a semantic space like `word2vec` is semantics preserving, e.g., the semantic topology is well-defined in such a representation space ([17]).

The semantic encoder is defined as a series of fully-connected layers to project the seen class prototypes $\mathcal{Z}^s$ onto the shared feature space.

**c) Cross-Modal Decoders:** We use decoder branches $g_{AB}(;\theta_{AB})$ and $g_{BA}(;,\theta_{BA})$ which is aimed to reconstruct the instances of $\mathcal{A}$ given the latent representations $f_B(\mathcal{B}^s)$ corresponding to modality $\mathcal{B}$ and vice-versa. We find that the decoder module acts as a regularizer and helps in mitigating any domain difference between the multi-modal data in the latent feature space.

### 3.2. Training

For stage 1, we fine-tune the pre-trained CNN models preferably with a lower learning rate on $(\mathcal{A}^s, \mathcal{Y}^s)$ and $(\mathcal{B}^s, \mathcal{Y}^s)$ and fine-tune it for the dataset using a cross-entropy loss, for obtaining the initialized $f_A()$ and $f_B()$, respectively.

For stage 2 training, we consider the following component loss measures:

- Cross-Modal latent loss: ($\mathcal{L}_{cmd}$).
- Cross-Modal Triplet loss: ($\mathcal{L}_{3tl}$).
- Decoder losses: ($\mathcal{L}_{rcs}$).
- Classification loss: ($\mathcal{L}_{class}$).

We denote an instance of class $c$ as $(\mathcal{A}/\mathcal{B})^s_c$. For an instance of any class other than $c$, we use the notation $(\tilde{\mathcal{A}}/\tilde{\mathcal{B}})^s_c$. These losses and their contributions are individually explained in the following.

**Cross-modal latent loss ($\mathcal{L}_{cmd}$):** To reduce the discrepancies between shared representations for a pair $(a^s_i, b^s_i)$ having a common semantic representation $z^s$, we seek to reduce their mean-square error between the respective embeddings. In this light, we bring closer $f_A(a^s_i)$ and $f_B(b^s_i)$, by reducing their distances from $z^s$. This, in effect, reduces the cross-modal intra-class variance substantially. In particular, this loss term is defined as follows:

$$\mathcal{L}_{cmd} = ||f_A(\mathcal{A}^s_c) - f_Z(\mathcal{Z}^s_c)||^2_{\mathbf{F}} + ||f_B(\mathcal{B}^s_c) - f_Z(\mathcal{Z}^s_c)||^2_{\mathbf{F}} \tag{1}$$

**F** represents the Frobenious norm of a matrix.

**Cross-modal triplet loss ($\mathcal{L}_{3lt}$):** It is found that reconstruction loss alone is not enough to yield good results. By adding a triplet loss function, we can reduce the intra-class distances, and increase the inter-class distances by pushing the samples far apart in the embedding space. To incorporate the triplet loss in a cross-modal framework, we form two types of triads. The first set of triads is constructed by taking a sketch sample as an anchor, along with its corresponding class image instance, and a different class image instance. The second set of triads is constructed by taking an image as the anchor, along with its corresponding sketch instance from the same class and one sketch instance from a different class. This function also acts as a regularizer (shown in equation 2). Taking $d()$ as the Euclidean distance between two vectors, we define the loss function as ($\mathcal{L}_{3lt}$):

$$\mathcal{L}_{si} = \max\left(d(\mathrm{f}_A(\mathcal{A}^s_c), \mathrm{f}_B(\mathcal{B}^s_c)) - d(\mathrm{f}_A(\mathcal{A}^s_c), \mathrm{f}_B\left(\tilde{\mathcal{B}}^s_c\right)) + \alpha, 0\right) \tag{2}$$

$$\mathcal{L}_{is} = \max\left(d(\mathrm{f}_B(\mathcal{B}^s_c), \mathrm{f}_A(\mathcal{A}^s_c)) - d(\mathrm{f}_B(\mathcal{B}^s_c), \mathrm{f}_A\left(\tilde{\mathcal{A}}^s_c\right)) + \alpha, 0\right) \tag{3}$$

$$\mathcal{L}_{3lt} = \mathcal{L}_{is} + \mathcal{L}_{si} \tag{4}$$

**Decoder loss ($\mathcal{L}_{rcs}$):** We note that **H** should achieve domain-independence by reducing any distributions-gap between $\mathcal{A}$ and $\mathcal{B}$. To enforce the domain-invariance, we encourage cross-domain sample reconstruction. In particular, for a given $(a^s_i, b^s_i, l^s_i)$, we aim for reconstructing $b^s_i$ from $f_A(a^s_i)$ and vice-versa. Since the sketch and image data drastically vary in appearance, such a cross-modal reconstruction regularizer helps in better alignment of the cross-modal data class-wise in **H**. The respective loss term is defined as:

$$\mathcal{L}_{rcs} = ||g_{AB}(f_A(\mathcal{A}^s_c)) - f_B(\mathcal{B}^s_c)||^2_{\mathbf{F}} + ||g_{BA}(f_B(\mathcal{B}^s_c)) - f_A(\mathcal{A}^s_c)||^2_{\mathbf{F}} \tag{5}$$

**Classification loss ($\mathcal{L}_{class}$):** To preserve the class information in the shared space, we use a cross-entropy loss function and learn the network parameters. We use this to preserve the class label information of $\mathcal{A}^s$ and $\mathcal{B}^s$. Hence, the classification loss is defined as:

$$\mathcal{L}_{class} = \text{CE}\left(f_A(\mathcal{A}^s)\right) + \text{CE}\left(f_B(\mathcal{B}^s)\right) \tag{6}$$

**Algorithm 1** The proposed training and inference stage

**Input:** $\mathcal{S} = \{\mathcal{A}^s, \mathcal{B}^s, \mathcal{Y}^s, \mathcal{Z}^s\}$
**Output:** Unified representations $\mathbf{H}$.

1: **Stage 1:** Normalize and pre-train $\mathcal{A}^s$ and $\mathcal{B}^s$.
2: **Stage 2:** Find the Word2Vec embeddings of $\mathcal{Y}^s$.
3: Train the network to obtain $\mathbf{H}$ by optimizing $\mathcal{L}$.
4: **do**
5:

$$\min_{\theta_A, \theta_B, \theta_{AB}, \theta_{BA}} \mathcal{L}_{cmd} + \mathcal{L}_{3lt} + \mathcal{L}_{class} + \mathcal{L}_{rcs} \tag{7}$$

6: Train $Z$
7: **while** until convergence
8: **return** $\theta_A, \theta_B, \theta_{AB}, \theta_{BA}$ (to realize $\mathbf{H}$)

**Input:** $a_i^u \in \mathcal{A}^u$ or $b_i^u \in \mathcal{B}^u$ and $\mathcal{Z}^u$
**Output:** Top-$K$ retrieved data.

9: Cross-modal zero-shot retrieval using $k$-NN.

### 3.3. Overall Objective function ( $\mathcal{L}$):

The final objective function is given by a linear combination of sum of the above-mentioned losses (shown in equation 8). Therefore, the overall objective function can be represented as:

$$\mathcal{L} = \mathcal{L}_{cmd} + \mathcal{L}_{rcs} + \mathcal{L}_{3lt} + \mathcal{L}_{class} \tag{8}$$

Since we have multiple losses, our problem becomes a non-convex optimization problem. To solve this, we reduce each loss term individually, keeping the other losses constant. The problem transforms into a convex optimization problem for that loss. We use an iterative shrinkage-based stochastic mini-batch gradient descent procedure to train the framework. Once the mapping of the data instances is made on the shared embedding space, the visual samples become closer to their corresponding semantic vectors. Algorithm 1 shows the overall approach for training the network.

### 3.4. Cross-modal Retrieval

For the retrieval stage, once the network has been fully trained, we save the trained network and save the weights for unseen classes instances of both the modalities. Using the feature embeddings of these instances, we choose a random query sample and find the $k$-nearest neighbor distance (Euclidean distance) from all other remaining samples. The top-$k$ retrieved images are considered as the retrieved results (shown in equation 9). This process is followed for both the uni-modal and cross-modal retrieval. The $\mathbf{D}$ matrix, as obtained from equation 3.4 is sorted to find the top-$k$ matches.

$$\mathbf{D} = ||\mathbf{H}_{\mathcal{A}/\mathcal{B}^{\mathbf{u}}} - \mathbf{H}_{\mathcal{A}/\mathcal{B}^{\mathbf{u}}}(\mathbf{q})||_{\mathbf{2}}^{\mathbf{2}} \tag{9}$$

Here $q$ represents the query image. The idea is to design such a robust shared embedding space that even a simple $l_2$-norm is enough for retrieval.

## 4. Experiments

In this section, we discuss in detail the training and the experimental protocols that have been used.

**Datasets:** We perform our experiments on two standard benchmark SBIR datasets, namely the Sketchy [20] and the TU-Berlin [13] dataset.

The Sketchy dataset [20] is a large scale dataset of sketch-photo pairs. Each photo consists of multiple corresponding sketches. There are 125 classes. Each class consists of 100 photos, and a variable number of sketches. Therefore the total number of photos is 12,500, while the number of sketches is 75,471. For our experiments, we choose randomly 25 classes for the testing phase as the unseen classes ($\mathcal{U}$). The remaining 100 classes were used for the training phase as the seen classes ($\mathcal{S}$).

The TU-Berlin dataset [13] consists of 250 classes, with 20,000 sketches and 204,489 photos. We train the network on 220 classes and test the results from the remaining 30 classes.

**Model Architecture:** To pre-train the network to get sufficiently discriminative initialization weights, we use a pre-trained ResNet50 and a VGG-16 based transfer-learning from the sketch and the image data. The network was fine-tuned to enhance its effectiveness on our dataset. We get a resultant feature vector of size 2048-d. The network was trained using a momentum optimizer, and a learning rate of 0.001 was chosen. To design the second stage, we extract the 300-d Word2Vec embedding of the labels of the datasets.

In the encoder part, we have kept three layers of fully-connected layer for $f_A()$ and $f_B()$. Similarly, for the decoder part, we have used a single fully-connected layer after the last layer of the encoders for $f_{AB}()$ and $f_{BA}()$. We have also added a batch normalization layer and induced a non-linearity function at the output of every fully-connected layer using *ReLU()*. For our experiments, we chose a learning rate of 0.001 and optimized the loss function using the Adam optimizer with a stochastic mini-batch gradient de-

Table 1: Performance of the proposed ZSCMR framework on the **Sketchy** and **TU-Berlin** dataset in terms of mAP and precision at top-100 (P@100) values and their corresponding embedding vector dimensions (Sketch→Image).

| | Task | Sketchy | | TU-Berlin | | size |
|---|---|---|---|---|---|---|
| | | **mAP** | **P@100** | **MAP** | **P@100** | |
| **SBIR** | Siamese CNN [18] | 0.183 | 0.143 | 0.153 | 0.122 | 64 |
| | SaN [26] | 0.129 | 0.104 | 0.112 | 0.096 | 512 |
| | 3D Shape [24] | 0.070 | 0.062 | 0.063 | 0.057 | 64 |
| | DSH (Binary) [13] | 0.171 | 0.231 | 0.129 | 0.189 | 64 |
| | GDH (Binary) [27] | 0.187 | 0.295 | 0.135 | 0.212 | 64 |
| | GN Triplet [20] | 0.204 | 0.296 | 0.175 | 0.253 | 1024 |
| **ZSL** | SSE [28] | 0.154 | 0.108 | 0.133 | 0.096 | 100 |
| | JLSE [29] | 0.131 | 0.185 | 0.109 | 0.155 | 220 |
| | ZSH [25] | 0.159 | 0.214 | 0.141 | 0.177 | 64 |
| | SAE [12] | 0.216 | 0.293 | 0.167 | 0.221 | 300 |
| **ZSL:SBIR** | ZS-SBIR [11] | 0.196 | 0.284 | 0.005 | 0.001 | 1024 |
| | ZSIH (Binary) [21] | 0.258 | 0.342 | 0.223 | 0.294 | 64 |
| | EMS [15] | - | - | 0.259 | 0.369 | 512 |
| | EMS (Binary) [15] | - | - | 0.165 | 0.252 | 64 |
| | CAAE [11] | 0.196 | 0.284 | - | - | 4096 |
| | CVAE [11] | 0.225 | 0.333 | - | - | 4096 |
| | SEM-PCYC [5] | 0.349 | 0.463 | 0.297 | 0.426 | 64 |
| | SAKE [14] | 0.364 | 0.487 | 0.359 | **0.481** | 64 |
| | **ZSCMR** | **0.467** | **0.510** | **0.362** | 0.429 | 64 |

Table 2: Performance of the proposed ZSCMR framework on the Sketchy and TU-Berlin datasets in terms of mAP and precision at top-100 (P@100) values.

| **Task** | **Sketchy** | | **TU-Berlin** | |
|---|---|---|---|---|
| | **mAP** | **P@100** | **MAP** | **P@100** |
| Sketch→Image | 0.467 | 0.510 | 0.362 | 0.429 |
| Image→Sketch | 0.429 | 0.451 | 0.298 | 0.353 |
| Sketch→Sketch | 0.444 | 0.559 | 0.318 | 0.397 |
| Image→Image | 0.686 | 0.718 | 0.605 | 0.653 |

scent approach. We trained the model with a batch size of 128, for 40 epochs. The $\alpha$ value of the triplet loss was heuristically set to 1.

**Training and Evaluation Protocol:** This network is then followed by a series of three fully connected networks, responsible for obtaining the shared embedding space for all the three the modalities of data. The triplets were selected by taking random anchor images from the training dataset and taking the same class and a different class sample from the other modality. The number of anchors from each modality were kept the same (10,000 triplets each). For the evaluation of our results, we use the standard mAP (mean average precision) and P@100 (precision for top-100) scores. Also, we compare the performance of our method on the Sketchy and the TU-Berlin dataset to the state-of-the-art algorithm [5]. A few more comparisons have been shown with [13, 27, 20, 29, 25, 12, 11, 21, 14].

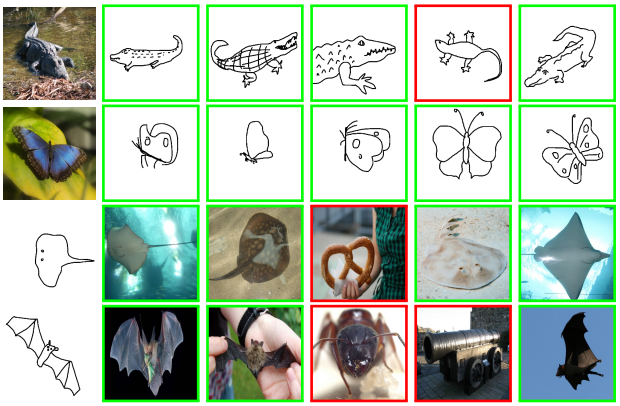

Figure 2. Top retrieved results of zero-shot cross-modal retrieval. Alternate rows represent Sketch→Image and Image→Sketch retrievals.

## 5. Results and Discussions

The train and test classes were chosen randomly for the experiments to avoid any bias induced while training. We report the performance of our model for both cross-modal and uni-modal retrieval in table 2.

Table 1 shows our and the other state-of-the-art model. The comparative study has been divided into three subparts. The first part shows the retrieval results using a simple SBIR framework, and the second part shows the results using ZSL, and the third set shows the algorithms which

have both ZSL and an SBIR framework. It can be seen that not only our framework beats state-of-the-art results with a high margin, it is also simultaneously capable of encoding the cross-modal and uni-modal neighborhood knowledge. Fig. 2 shows the top retrieved images, given a query image (first column). The images with green boundaries represent the correctly retrieved classes, while the red ones represent the incorrectly retrieved images.

### 5.1. Ablation Studies

To investigate the contribution of each of the losses in our system, we perform an ablation study wherein we train the network, excluding one loss at a time. This helps us to understand the contribution of each loss in the proposed network. We perform the ablation studies for both the Sketchy and the TU Berlin datasets and compare their corresponding SBIR and Image-based sketch retrieval (IBSR) performances. Figure 3 shows the bar graph of the performance of each of these studies in terms of mAP values.

For the first set of experiments, we keep all losses as given in the objective function in equation 8. The results obtained from this has been reported and explained in the previous sections. Next, we trained the model using equation 8, leaving the cross-modal latent loss. The performance falls drastically as the feature embedding of both the modalities fall quite apart from each other.

For the third set of experiments, we omit the cross-entropy classifier loss from the overall objective function. As seen from figure 3, the performance of the model drops even more, as, in this case, the class information embedding in the shared space is lost. The shared latent space is no longer sufficiently discriminative, and there is a considerable reduction in the intra-class distances.

For the next set of experiments, we drop the cross-triplet loss function from the total loss equation. We can see a rise in the performance of the model as compared to the previous experiments. This loss helps in bringing down the inter-modal distances while pushing the samples from the same classes further apart in the feature space. Finally, we look at the effect of the network without the decoder loss function. This provides an immense boost in the performance of the model in combination with the triplet loss. It helps the network achieve domain Independence. The exact effect is illustrated in figure 3.

It can be seen from the bar graphs that while the triplet loss and decoder losses help in beating the state-of-the-art results, the cross-latent loss and the cross-entropy losses are the essential loss functions without which the performance of the network fails miserably.

## 6. Conclusion

We propose a cross-modal zero-shot retrieval framework and evaluate the results on sketch-based data. The implementation of our project can be found from the GitHub link: https://github.com/ushasi/A-Simplified-framework-for-Zero-shot-Cross-Modal-Sketch-Data-Retrieval. The main motive of our problem statement is to project different domain data onto a common embedding space, wherein discrimination between different classes within different data can be effectively possible. The proposed framework not only beats the current state-of-the-art results in unseen classes SBIR, but it also successfully encodes the image-based sketch retrieval, as well as the two uni-modal retrieval data information in the unseen classes. This has been made possible by making different classes separable in the semantic space. We are currently interested in exploring further in this domain and investigating the results with incremental learning.

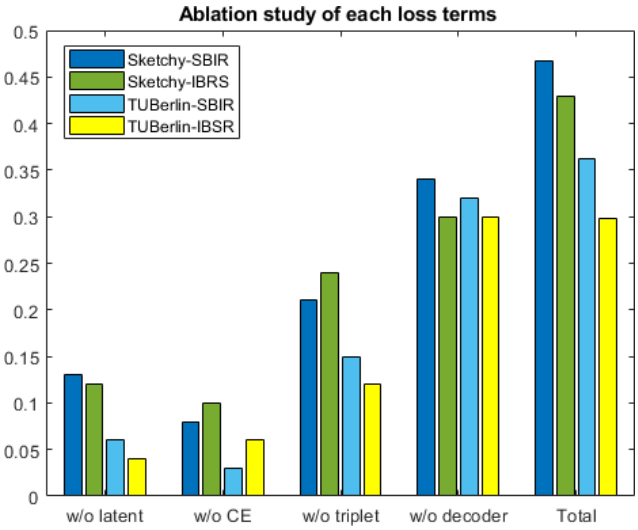


Figure 3. Ablation study of each of the loss terms in the Sketchy and the TU-Berlin dataset. The bar-graph shows both sketch-based image retrieval and image-based sketch retrieval performances, in terms of mAP values.

## References


[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 2

[2] Ushasi Chaudhuri, Biplab Banerjee, and Avik Bhattacharya. Siamese graph convolutional network for content based remote sensing image retrieval. *Computer Vision and Image Understanding*, 184:22–30, 2019. 2

[3] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu. Cmir-net: A deep learning based model for cross-modal retrieval in remote sensing. *arXiv preprint arXiv:1904.04794*, 2019. 2

[4] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 415–423, 2015. 2

[5] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2019. 2, 6

[6] Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa. Retrieval of difficult image classes using svd-based relevance feedback. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 23–30. ACM, 2004. 2

[7] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018. 2

[8] Inge Gavat, Daniela Faur, Marius I Piso, Florin Serban, and Mihai Datcu. Knowledge based image information mining system-kim used for flooding and other risk assessments. In *PV 2007 International Conference on Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data*. 2

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[10] Omkar Gune, Biplab Banerjee, and Subhasis Chaudhuri. Structure aligning discriminative latent embedding for zero-shot learning. In *BMVC*, page 218, 2018. 2

[11] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018. 2, 6

[12] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017. 2, 6

[13] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2871, 2017. 5, 6

[14] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3662–3671, 2019. 2, 6

[15] Peng Lu, Gao Huang, Yanwei Fu, Guodong Guo, and Hangyu Lin. Learning large euclidean margin for sketch-based image retrieval. *arXiv preprint arXiv:1812.04275*, 2018. 6

[16] Devraj Mandal, Kunal N Chaudhury, and Soma Biswas. Generalized semantic preserving hashing for n-label cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4076–4084, 2017. 2

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 4

[18] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2460–2464. IEEE, 2016. 6

[19] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016. 2

[20] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016. 5, 6

[21] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3598–3607, 2018. 2, 6

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[23] Antonio Torralba and Aude Oliva. Depth estimation from image structure. *IEEE Transactions on pattern analysis and machine intelligence*, 24(9):1226–1238, 2002. 2

[24] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1883, 2015. 6

[25] Zhilin Yang, William Cohen, and Salakhutdinov Ruslan. Revisiting semi-supervised learning with graph embeddings. *arXiv preprint arXiv:1603.08861*, 2016. 2, 6

[26] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425, 2017. 6

[27] J Zhang, F Shen, L Liu, F Zhu, M Yu, L Shao, H Tao Shen, and L Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 297–314, 2018. 2, 6

[28] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015. 6

[29] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016. 2, 6