Zero-Shot Sketch-Image Hashing

Yuming Shen* [1], Li Liu*[2], Fumin Shen [3], and Ling Shao [2,1]

[1]School of Computing Sciences, University of East Anglia, Norwich, UK
[2]Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
[3]Center for Future Media, University of Electronic Science and Technology of China, China
yuming.shen@uea.ac.uk, liuli1213@gmail.com, fumin.shen@gmail.com, ling.shao@ieee.org

Abstract

Recent studies show that large-scale sketch-based image retrieval (SBIR) can be efficiently tackled by cross-modal binary representation learning methods, where Hamming distance matching significantly speeds up the process of similarity search. Providing training and test data subjected to a fixed set of pre-defined categories, the cutting-edge SBIR and cross-modal hashing works obtain acceptable retrieval performance. However, most of the existing methods fail when the categories of query sketches have never been seen during training.

In this paper, the above problem is briefed as a novel but realistic zero-shot SBIR hashing task. We elaborate the challenges of this special task and accordingly propose a zero-shot sketch-image hashing (ZSIH) model. An end-to-end three-network architecture is built, two of which are treated as the binary encoders. The third network mitigates the sketch-image heterogeneity and enhances the semantic relations among data by utilizing the Kronecker fusion layer and graph convolution, respectively. As an important part of ZSIH, we formulate a generative hashing scheme in reconstructing semantic knowledge representations for zero-shot retrieval. To the best of our knowledge, ZSIH is the first zero-shot hashing work suitable for SBIR and cross-modal search. Comprehensive experiments are conducted on two extended datasets, i.e., Sketchy and TU-Berlin with a novel zero-shot train-test split. The proposed model remarkably outperforms related works.

1. Introduction

Matching real images with hand-free sketches has recently aroused extensive research interest in computer vision, multimedia and machine learning, forming the term of sketch-based image retrieval (SBIR). Differing the conventional text-image cross-modal retrieval, SBIR delivers a more applicable scenario where the targeted candidate images are conceptually unintelligible but visualizable to user. Several works have been proposed handling the SBIR task by learning real-valued representations [16, 17, 23, 24, 49, 50, 52, 57, 59, 69, 70]. As an extension of conventional data hashing techniques [51, 20, 39, 54, 53], cross-modal hashing [4, 13, 72, 36, 33, 26, 5, 6] show great potential in retrieving heterogeneous data with high efficiency due to the computationally cheap Hamming space matching, which is recently adopted to large-scale SBIR in [38] with impressive performance. Entering the era of big data, it is always feasible and appreciated to seek binary representation learning methods for fast SBIR.

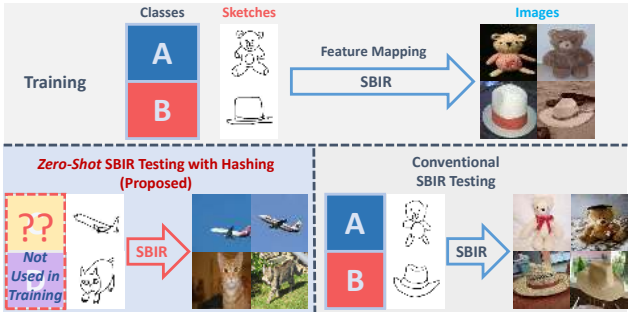

Figure 1. In conventional SBIR and cross-modal hashing (**bottom right**), the categories of training data include the ones of test data, marked as *'A'* and *'B'*. For our *zero-shot* task (**bottom left**), training data are still subjected to class *'A'* and *'B'*, but test sketches and images are coming from other categories, *i.e.*, *'plane'* and *'cat'* in this case. Note that data labels are not used as test inputs and the test data categories shall be unknown to the learning system.

However, the aforementioned works suffer from obvious drawbacks. Given a fixed set of categories of training and test data, these methods successfully manage to achieve sound SBIR performance, which is believed to be a relatively easy task as the visual knowledge from all concepts has been explored during parameter learning, while

*Yuming Shen and Li Liu contributed equally to this work.

in a real-life scenario, there is no guarantee that the training data categories cover all concepts of potential retrieval queries and candidates in the database. An extreme case occurs when test data are subjected to an absolutely different set of classes, excluding the trained categories. Unfortunately, experiments show that existing cross-modal hashing and SBIR works generally fail on this occasion as the learned retrieval model has no conceptual knowledge about what to find.

Considering both the **train-test category exclusion** and **retrieval efficiency**, a novel but realistic task yields *zero-shot* SBIR hashing. Fig. 1 briefly illustrates the difference between our task and conventional SBIR task. In conventional SBIR and cross-modal hashing, the categories of training data include the ones of test data, marked as *'A'* and *'B'* in Fig. 1. On the other hand, for the *zero-shot* task, though training data are still subjected to class *'A'* and *'B'*, test sketches and images are coming from other categories, *i.e.*, *'plane'* and *'cat'* in this case. In the rest of this paper, we denote the training and test categories as *seen* and *unseen* classes, since they are respectively known and unknown to the retrieval model.

Our *zero-shot* SBIR hashing setting is a special case of *zero-shot* learning in inferring knowledge out of the training samples. However, existing works basically focus on single-modal *zero-shot* recognition [56, 75, 76, 31], and are not suitable for efficient image retrieval. In [67], an inspiring *zero-shot* hashing scheme is proposed for large-scale data retrieval. Although [67] suggests a reasonable *zero-shot* train-test split close to Fig. 1 for retrieval experiments, it is still not capable for cross-modal hashing and SBIR.

Regarding the drawbacks and the challenging task discussed above, a novel *zero-shot* sketch-image hashing (ZSIH) model is proposed in this paper, simultaneously delivering (1) cross-modal hashing, (2) SBIR and (3) *zero-shot* learning. Leveraging state-of-the-art deep learning and generative hashing techniques, we formulate our deep network according to the following problems and themes:

(a) Not all regions in an image or sketch are informative for cross-modal mapping.

(b) The heterogeneity between image and sketch data needs to be mitigated during training to produce unified binary codes for matching.

(c) Since visual knowledge alone is inadequate for *zero-shot* SBIR hashing, a back-propagatable deep hashing solution transferring semantic knowledge to the *unseen* classes is desirable.

The contributions of this work are summarized as follows:

- To the best of our knowledge, ZSIH is the first *zero-shot* hashing work for large-scale SBIR.
- We propose an end-to-end three-network structure for deep generative hashing, handling the train-test category exclusion and search efficiency with attention model, Kronecker fusion and graph convolution.
- The ZSIH model successfully produces reasonable retrieval performance under the *zero-shot* setting, while existing methods generally fail.

Related Works. General cross-modal binary representation learning methods [4, 13, 72, 33, 58, 43, 36, 26, 6, 63, 18, 55, 66, 5, 42] target to map large-scale heterogeneous data with low computational cost. SBIR, including fine-grained SBIR, learns shared representations to specifically mitigate the expressional gap between hand-crafted sketches and real images [16, 17, 23, 24, 47, 48, 49, 50, 52, 57, 59, 62, 69, 70, 77, 78], while the efficiency issue is not considered. *Zero-shot* learning [19, 31, 75, 76, 56, 46, 7, 2, 64, 3, 34, 11, 8, 25, 74, 14, 35, 65, 27, 68, 40] is also related to our work, though it does not originally focus on cross-modal retrieval. Among the existing researches, zero-shot hashing (ZSH) [67] and deep sketch hashing (DSH) [38] are the two closest works to this paper. DSH [38] considers fast SBIR with deep hashing technique, but it fails to handle the *zero-shot* setting. ZSH [67] extends the traditional *zero-shot* task to a retrieval scheme.

2. The Proposed ZSIH Model

This work focuses on solving the problem of hand-free SBIR using deep binary codes under the *zero-shot* setting, where the image and sketch data belonging to the *seen* categories are only used for training. The proposed deep networks are expected to be capable for encoding and matching the *unseen* sketches with images, categories of which have never appeared during training.

We consider a multi-modal data collection $\mathcal{O}^c = \{\mathbf{X}^c, \mathbf{Y}^c\}$ from *seen* categories $\mathcal{C}^c$ covering both real images $\mathbf{X}^c = \{\mathbf{x}_i^c\}_{i=1}^N$ and sketch images $\mathbf{Y}^c = \{\mathbf{y}_i^c\}_{i=1}^N$ for training, where N indicates the set size. For the simplicity of presentation, it is assumed that image and sketch data with the same index i, *i.e.*, $\mathbf{x}_i^c$ and $\mathbf{y}_i^c$ share the same category label. Additionally, similar to many conventional *zero-shot* learning algorithms, our model requires a set of semantic representations $\mathbf{S}^c = \{\mathbf{s}_i^c\}_{i=1}^N$ in transferring supervised knowledge to the *unseen* data. The aim is to learn two deep hashing functions $f(\cdot)$ and $g(\cdot)$ for images and sketches respectively. Given a set of image-sketch data $\mathcal{O}^u = \{\mathbf{X}^u, \mathbf{Y}^u\}$ belonging to the *unseen* categories $\mathcal{C}^u$ for test, the proposed deep hashing functions encode these *unseen* data into binary codes, *i.e.*, $f : \mathbb{R}^d \rightarrow \{0,1\}^M, g : \mathbb{R}^d \rightarrow \{0,1\}^M$, where d refers to the original data dimensionality and M is the targeted hash code length. Concretely, as the proposed model handles SBIR under the *zero-shot* setting, there should be no intersection between the *seen* categories for training and the *unseen* classes for test, *i.e.*, $\mathcal{C}^c \bigcap \mathcal{C}^u = \varnothing$.

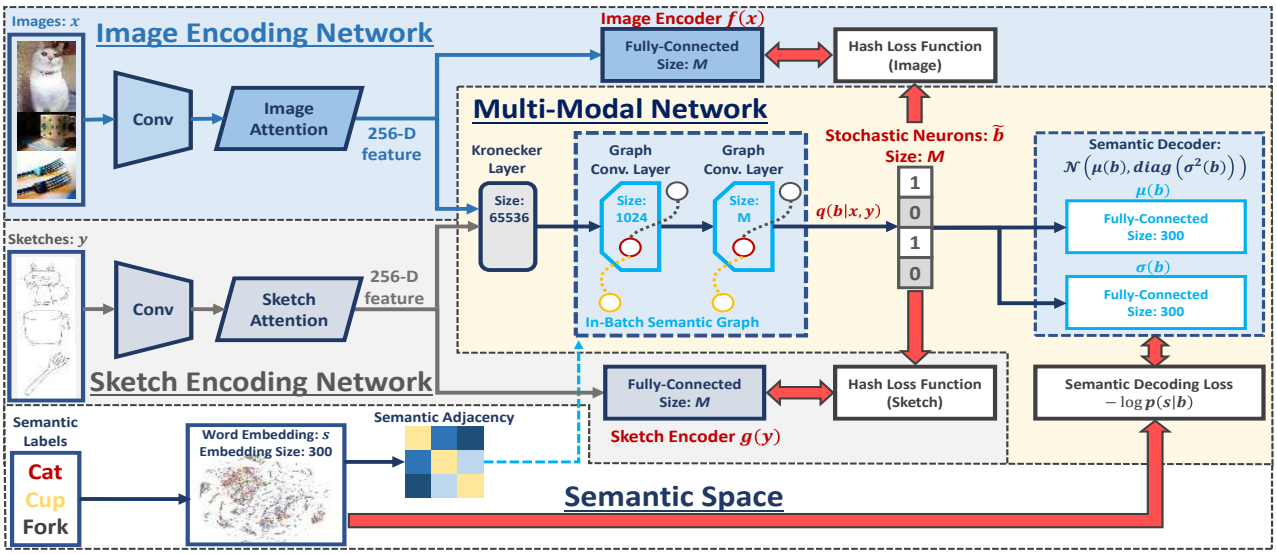

Figure 2. The deep network structure of ZSIH. The image (in light blue) and sketch encoding network (in grey) act as hash function for the respective modality with attention models [59]. The multi-modal network (in canary yellow) only functions during training. Sketch-image representations are fused by a Kronecker layer [22]. Graph convolution [30] and generative hashing techniques are leveraged to explore the semantic space for *zero-shot* SIBR hashing. **Network configurations** are also provided here.

2.1. Network overview

The proposed ZSIH model is an end-to-end deep neural network for *zero-shot* sketch-image hashing. The architecture of ZSIH is illustrated in Fig. 2, which is composed of three concatenated deep neural networks, *i.e.*, the image/sketch encoders and the multi-modal network, to tackle the problems discussed above.

2.1.1 Image/sketch encoding networks

As is shown in Fig. 2, the networks with light blue and grey background refer to the binary encoders $f(\cdot)$ and $g(\cdot)$ for images and sketches respectively. An image or sketch is firstly rendered to a set of corresponding convolutional layers to produce a feature map, and then the attention model mixes informative parts into a single feature vector for further operation. The AlexNet [32] before the last pooling layer is built to obtain the feature map. We introduce the attention mechanism in solving issue (a), of which the structure is close to [59] with weighted pooling to produce a 256-D feature. Binary encoding is performed by a fully-connected layer taking input from the attention model with a `sigmoid` nonlinearity. During training, $f(\cdot)$ and $g(\cdot)$ are regularized by the output of the multi-modal network, so these two encoders are supposed to be able learn modal-free representations for *zero-shot* sketch-image matching.

2.1.2 Multi-modal network as code learner

The multi-modal network only functions during training. It learns the joint representations for sketch-image hashing, handling the problem (b) of modal heterogeneity. One possible solution for this is to introduce a fused representation layer taking inputs from both image and sketch modality for further encoding. Inspired by Hu *et al.* [22], we find the Kronecker product fusion layer suitable for our model, which is discussed in Sec. 2.2. Shown in Fig. 2, the Kronecker layer takes inputs from the image and sketch attention model, and produces a single feature vector for each pair of data points. We index the training images and sketches in a coherent category order. Therefore the proposed network is able to learn compact codes for both images and sketches with clear categorical information.

However, simply mitigating the model heterogeneity does not fully solves the challenges in ZSIH. As is mentioned in problem (c), for *zero-shot* tasks, it is essential to leverage the semantic information of training data to generalize knowledge from the *seen* categories to the *unseen* ones. Suggested by many *zero-shot* learning works [31, 19, 67], the semantic representations, *e.g.*, word vectors [44], implicitly determine the category-level relations between data points from different classes. Based on this, during the joint code learning process, we novelly enhance the hidden neural representations by the semantic relations within a batch of training data using the graph convolutional networks (GCNs) [10, 30]. It can be observed in Fig. 2 that two graph convolutional layers are built in the multi-modal network, successively following the Kronecker layer. In this way, the in-batch data points with strong latent semantic relations are entitled to interact during gradient computation. Note that the output length of the second

graph convolutional layer for each data point is exactly the target hash code length, *i.e.*, M. The formulation of the semantic graph convolution layer is given in Sec. 2.3.

To obtain binary codes as the supervision of $f(\cdot)$ and $g(\cdot)$, we introduce the stochastic generative model [9] for hashing. A back-propagatable structure of stochastic neurons is built on the top of the second graph convolutional layer, producing hash codes. Shown in Fig. 2, a decoding model is topped on the stochastic neurons, reconstructing the semantic information. By maximizing the decoding likelihood with gradient-based methods, the whole network is able to learn semantic-aware hash codes, which also accords our perspective of issue (c) for *zero-shot* sketch-image hashing. We elaborate on this design in Sec. 2.4 and 2.5.

2.2. Fusing sketch and image with Kronecker layer

Sketch-image feature fusion plays an important role in our task as is addressed in problem (b) of Sec. 1. An information-rich fused neural representation is in demand for accurate encoding and decoding. To this end, we utilize the recent advances in Kronecker-product-based feature learning [22] as the fusion network. Denoting the attention model outputs of a sketch-image pair $\{\mathbf{y}, \mathbf{x}\}$ from the same category as $\mathbf{h}^{(sk)} \in \mathbb{R}^{256}$ and $\mathbf{h}^{(im)} \in \mathbb{R}^{256}$, a non-linear data fusion operation can be derived as

$$\mathcal{W} \times_1 \mathbf{h}^{(sk)} \times_3 \mathbf{h}^{(im)}. \tag{1}$$

Here $\mathcal{W}$ is a third-order tensor of fusion parameters and $\times$ denotes tensor dot product. We use the left subscript to indicate on which axis tensor dot operates. De-compositing $\mathcal{W}$ with Tucker decomposition [61], the fused output of the Kronecker layer $\mathbf{h}^{(kron)}$ in our model is derived as

$$\mathbf{h}^{(kron)} = \delta\big((\mathbf{h}^{(sk)}\mathbf{W}^{(sk)}) \otimes (\mathbf{h}^{(im)}\mathbf{W}^{(im)})\big), \tag{2}$$

resulting in a 65536-D feature vector. Here $\otimes$ is the Kronecker product operation between two tensors, and $\mathbf{W}^{(sk)}, \mathbf{W}^{(im)} \in \mathbb{R}^{256\times256}$ are trainable linear transformation parameters. $\delta(\cdot)$ refers to the activation function, which is the `ReLU` [45] nonlinearity for this layer.

Kronecker layer [22] is supposed to be a better choice in feature fusion for ZSIH than many conventional methods such as layer concatenation or factorized model [71]. This is because the Kronecker layer largely expands the feature dimensionality of the hidden states with a limited number of parameters, and thus consequently stores more expressive structural relation between sketches and images.

2.3. Semantic-relation-enhanced hidden representation with graph convolution

In this subsection, we describe how the categorical semantic relations are enhanced in our ZSIH model using GCNs. Considering a batch of training data $\{\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i\}_{i=1}^{N_B}$ consisting of N_B category-coherent sketch-image pairs with their semantic representations $\{\mathbf{s}_i\}$, we denote the hidden state of the l-th layer in the multi-modal network of this training batch as $\mathbf{H}^l$ to be rendered to a graph convolutional layer. As is mentioned in Sec. 2.1.2, for our graph convolutional layers, each training batch is regarded as an N_B-vertex graph. Therefore, a convolutional filter g_θ parameterized by θ can be applied to $\mathbf{H}^l$, producing the $(l+1)$-th hidden state $\mathbf{H}^{(l+1)} = g_\theta * \mathbf{H}^{(l)}$. Suggested by [30], this can be approached by a layer-wise propagation rule, *i.e.*,

$$\mathbf{H}^{(l+1)} = \delta\big(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}_\theta\big), \tag{3}$$

using the first-order approximation of the localized graph filter [10, 21]. Again, here $\delta(\cdot)$ is the activation function and $\mathbf{W}_\theta$ refers to the linear transformation parameter. $\mathbf{A}$ is an $N_B \times N_B$ self-connected in-batch adjacency and $\mathbf{D}$ can be defined by $\mathbf{D} = \mathtt{diag}(\mathbf{A}\mathbb{1})$. It can be seen in Fig. 2 that the in-batch adjacency $\mathbf{A}$ is determined by the semantic representations $\{\mathbf{s}_i\}$, of which each entry $\mathbf{A}^{(j,k)}$ can be computed by $\mathbf{A}^{(j,k)} = e^{-\frac{\|\mathbf{s}_j - \mathbf{s}_k\|^2}{t}}$. In the proposed ZSIH model, two graph convolutional layers are built, with output feature dimensions of $N_B \times 1024$ and $N_B \times M$ for a whole batch. We choose the `ReLU` nonlinearity for the first layer and the `sigmoid` function for the second one to restrict the output values between 0 and 1.

Intuitively, the graph convolutional layer proposed by [30] can be construed as performing elementary row transformation on a batch of data from fully-connected layer before activation according to the graph Laplacian of $\mathbf{A}$. In this way, the semantic relations between different data points are intensified within the network hidden states, benefiting our *zero-shot* hashing model in exploring the semantic knowledge. Traditionally, correlating different deep representations can be tackled by adding a trace-like regularization term in the learning objective. However, this introduces additional hyper parameters to balance the loss terms and the hidden states in the network of different data points are still isolated.

2.4. Stochastic neurons and decoding network

The encoder-decoder model for ZSIH is introduced in this subsection. Inspired by [9], a set of latent probability variables $\mathbf{b} \in (0,1)^M$ are obtained from the second graph convolutional layer output respective to $\{\mathbf{x}, \mathbf{y}\}$ corresponding to the hash code for a sketch-image pair $\{\mathbf{x}, \mathbf{y}\}$ with the semantic feature $\mathbf{s}$. The stochastic neurons [9] are imposed to $\mathbf{b}$ to produce binary codes $\widetilde{\mathbf{b}} \in \{0,1\}^M$ through a sampling procedure:

$$\widetilde{\mathbf{b}}^{(m)} = \begin{cases} 1 & \mathbf{b}^{(m)} \geqslant \epsilon^{(m)}, \\ 0 & \mathbf{b}^{(m)} < \epsilon^{(m)}, \end{cases} \quad \text{for } m = 1 \dots M, \tag{4}$$

where $\epsilon^{(m)} \sim \mathcal{U}([0,1])$ are random variables. As is proved in [9], this structure can be differentiable, allowing error

Algorithm 1: The Training Procedure of ZSIH

Input: Sketch-image dataset $\mathcal{O} = \{\mathbf{X}, \mathbf{Y}\}$, semantic representations $\mathbf{S}$ and max training iteration T
Output: Network parameters Θ

repeat
  Get a random mini-batch $\{\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i\}_{i=1}^{N_B}$, assuring $\mathbf{x}_i, \mathbf{y}_i$ belong to the same class
  Build $\mathbf{A}$ according to semantic distances
  for $i = 1 \ldots N_B$ **do**
    Sample a set of $\epsilon^{(m)} \sim \mathcal{U}([0, 1])$
    Sample a set of $\widetilde{\mathbf{b}} \sim q(\mathbf{b}|\mathbf{x}_i, \mathbf{y}_i)$
  end
  $\mathcal{L} \leftarrow$ Eq. (7)
  $\Theta \leftarrow \Theta - \mathbf{\Gamma}(\nabla_\Theta \mathcal{L})$ according to Eq. (8)
until *convergence or max training iter T is reached*;

back-propagation from the decoder to the previous layers. Therefore, the posterior of $\mathbf{b}$, *i.e.*, $p(\mathbf{b}|\mathbf{x}, \mathbf{y})$, is approximated by a Multinoulli distribution:

$$q(\widetilde{\mathbf{b}}|\mathbf{x}, \mathbf{y}) = \prod_{m=1}^{M} (\mathbf{b}^{(m)})^{\widetilde{\mathbf{b}}^{(m)}} (1 - \mathbf{b}^{(m)})^{1-\widetilde{\mathbf{b}}^{(m)}}. \quad (5)$$

We follow the idea of generative hashing to build a decoder on the top of the stochastic neurons. During optimization of ZSIH, this decoder is regularized by the semantic representations $\mathbf{s}$ using the following Gaussian likelihood with the reparametrization trick [29], *i.e.*,

$$p(\mathbf{s}|\mathbf{b}) = \mathcal{N}\big(\mathbf{s}|\mu(\mathbf{b}), \mathtt{diag}(\sigma^2(\mathbf{b}))\big), \quad (6)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are implemented by fully-connected layers with identity activations. To this end, the whole network can be trained en-to-end. The learning objective is given in the next subsection.

2.5. Learning objective and optimization

The learning objective of the whole network for a batch of sketch and image data is defined as follows:

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{N_B} \mathbb{E}_{q(\mathbf{b}|\mathbf{x}_i, \mathbf{y}_i)} \big[\log q(\mathbf{b}|\mathbf{x}_i, \mathbf{y}_i) - \log p(\mathbf{s}_i|\mathbf{b}) \\ + \frac{1}{2M} \big(\|f(\mathbf{x}_i) - \mathbf{b}\|^2 + \|g(\mathbf{y}_i) - \mathbf{b}\|^2 \big) \big].\end{aligned} \quad (7)$$

Concretely, the expectation term $\mathbb{E}[\cdot]$ in Eq. (7) simulates the variational-like learning objectives [29, 9] as a generative model. However, we are not exactly lower-bounding any data prior distribution since it is generally not feasible for our ZSIH network. $\mathbb{E}[\cdot]$ here is an empirically-built loss, simultaneously maximizing the output code entropy via $\mathbb{E}_{q(\mathbf{b}|\mathbf{x},\mathbf{y})}[\log q(\mathbf{b}|\mathbf{x}, \mathbf{y})]$ and preserving the semantic knowledge for the *zero-shot* task by $\mathbb{E}_{q(\mathbf{b}|\mathbf{x},\mathbf{y})}[-\log p(\mathbf{s}|\mathbf{b})]$. The single-model encoding functions $f(\cdot)$ and $g(\cdot)$ are trained by the stochastic neurons outputs of the multi-modal network using L-2 losses. The sketch-image similarities can be reflected in assigning related sketches and images with the sample code. To this end, $f(\cdot)$ and $g(\cdot)$ are able to encode out-of-sample data without additional category information, as the imposed training codes are semantic-knowledge-aware. The gradient of our learning objective w.r.t. the network parameter Θ can be estimated by a Monte Carlo process in sampling $\widetilde{\mathbf{b}}$ using the small random signal ϵ according to Eq. (4), which can be derived as

$$\begin{aligned}\nabla_\Theta \mathcal{L} \simeq \sum_{i=1}^{N_B} \mathbb{E}_\epsilon \Big[\nabla_\Theta \Big(\log q(\widetilde{\mathbf{b}}|\mathbf{x}_i, \mathbf{y}_i) - \log p(\mathbf{s}_i|\widetilde{\mathbf{b}}) \\ + \frac{1}{2M} \big(\|f(\mathbf{x}_i) - \widetilde{\mathbf{b}}\|^2 + \|g(\mathbf{y}_i) - \widetilde{\mathbf{b}}\|^2 \big) \Big) \Big].\end{aligned} \quad (8)$$

As $\log q(\cdot)$ forms up into an inverse cross-entropy loss and $\log p(\cdot)$ is reparametrized, this estimated gradient can be easily computed. Alg. 1 illustrates the whole training process of the proposed ZSIH model, where the operator $\mathbf{\Gamma}(\cdot)$ refers to the Adam optimizer [28] for adaptive gradient scaling. Different from many existing deep cross-modal and *zero-shot* hashing models [5, 38, 67, 26] which require alternating optimization procedures, ZSIH can be efficiently and conveniently trained end-to-end with SGD.

2.6. Out-of-sample extension

When the network of ZSIH is trained, it is able to hash image and sketch data from the *unseen* classes $\mathcal{C}^u$ for matching. The codes can be obtained as follows:

$$\begin{aligned}\mathbf{B}^{im} &= (\mathrm{sign}(f(\mathbf{X}^u - 0.5)) + 1)/2 \in \{0, 1\}^{N^u \times M}, \\ \mathbf{B}^{sk} &= (\mathrm{sign}(g(\mathbf{Y}^u - 0.5)) + 1)/2 \in \{0, 1\}^{N^u \times M},\end{aligned} \quad (9)$$

where N^u is the size of test data. As is shown in Fig. 2, the encoding networks $f(\cdot)$ and $g(\cdot)$ are standing on their own. Semantic representations of test data are not required and there is no need to render data to the multi-modal network. Thus, encoding test data is non-trivial and can be efficient.

3. Experiments

3.1. Implementation details

The proposed ZSIH model is implemented with the popular deep learning toolbox Tensorflow [1]. We utilize the settings of AlexNet [32] pre-trained on ImageNet [12] before the last pooling layer to build our image and sketch CNNs. The attention mechanism is inspired by Song *et al.* [59] without the shortcut connection. The attended 256-D feature is obtained by a weighted pooling operation according to the attention map. All configurations of our network are provided in Fig. 2. We obtain the semantic representation of each data point using the 300-D word vector [44] according to the class name. When the class name is

Table 1. *zero-shot* SBIR mAP@all comparison between ZSIH and some cross-modal hashing baselines.

Method	Cross Modal	Binary Code	Zero Shot	Sketchy (Extended)			TU-Berlin (Extended)		
				32 bits	64 bits	128 bits	32 bits	64 bits	128 bits
ZSH [67]		✓	✓	0.146	0.165	0.168	0.132	0.139	0.153
CCA [60]	✓			0.092	0.089	0.084	0.083	0.074	0.062
CMSSH [4]	✓	✓		0.094	0.096	0.111	0.073	0.077	0.080
CMFH [13]	✓	✓		0.115	0.116	0.125	0.114	0.118	0.135
SCM-Orth [72]	✓	✓		0.105	0.107	0.093	0.089	0.092	0.095
SCM-Seq [72]	✓	✓		0.092	0.100	0.084	0.084	0.087	0.072
CVH [33]	✓	✓		0.076	0.075	0.072	0.065	0.061	0.055
SePH-Rand [36]	✓	✓		0.108	0.097	0.094	0.071	0.065	0.070
SePH-KM [36]	✓	✓		0.069	0.066	0.071	0.067	0.068	0.065
DSH [38]	✓	✓		0.137	0.164	0.165	0.119	0.122	0.146
ZSIH	✓	✓	✓	**0.232**	**0.254**	**0.259**	**0.201**	**0.220**	**0.234**

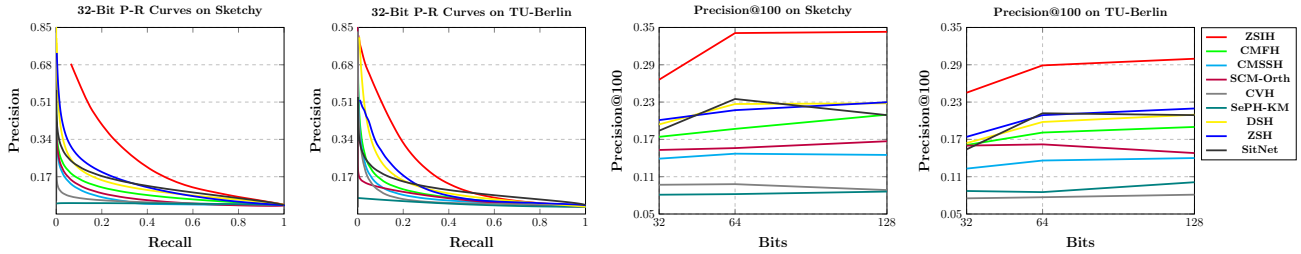

Figure 3. Precision-recall curves and precision@100 results of ZSIH and several hashing baselines are shown above. To keep the content concise, only 32-bit precision-recall curves are illustrated here.

not included in the word vector dictionary, it is replaced by a synonym. For all of our experiments, the hyper-parameter t is set to $t = 0.1$ with a training batch size of 250. Our network is able to be trained end-to-end.

3.2. Zero-shot experimental settings

To perform SBIR with binary codes under the novelly-defined *zero-shot* cross-modal setting, the experiments of this work are taken on two large-scale sketch datasets, *i.e.*, Sketchy [52] and TU-Berlin [15], with extended images obtained from [38]. We follow the SBIR evaluation metrics in [38] where sketch queries and image retrieval candidates with the same label are marked as relevant, while our retrieval performances are reported based on nearest neighbour search in the hamming space.

Sketchy Dataset [52] (Extended). This dataset originally consists of $75,471$ hand-drawn sketches and $12,500$ corresponding images from 125 categories. With the extended $60,502$ real images provided by Liu *et al.* [38], the total size of the whole image set yields $73,002$. We randomly pick 25 classes of sketches and images as the *unseen* test set for SBIR, and data from the rest 100 *seen* classes are used for training. During the test phase, the sketches from the *unseen* classes are taken as retrieval queries, while the retrieval gallery is built using all the images from the *unseen* categories. Note that the test classes are not presenting during training for *zero-shot* retrieval.

TU-Berlin Dataset [15] (Extended). The TU-Berlin dataset contains $20,000$ sketches subjected to 250 categories. We also utilize the extended nature images provided in [38, 73] with a total size of $204,489$. 30 classes of images and sketches are randomly selected to form the retrieval gallery and query set respectively. The rest data are used for training. Since the quantities of real images from different classes are extremely imbalanced, we additionally require each test category have at least 400 images when picking the test set.

3.3. Comparison with existing methods

As cross-modal hashing for SBIR under the *zero-shot* setting has never been proposed before to the best of our knowledge, the quantity of potential related existing baselines is limited. Our task can be regarded as a combination of conventional cross-modal hashing, SBIR and *zero-shot* learning. Therefore, we adopt existing methods according to these themes for retrieval performance evaluation. We use the *seen-unseen* splits identical to ours for training and testing the selected baselines. The deep-learning-based baselines are retrained end-to-end using the *zero-shot* setting mentioned above. For the non-deep baselines, we extract the respective AlexNet [32] `fc_7` features pre-trained on the *seen* sketches and images as model training inputs for a fair comparison with our deep model.

Cross-Modal Hashing Baselines. Several state-of-the-

Table 2. *Zero-shot* sketch-image retrieval performance comparison of ZSIH with existing SBIR and *zero-shot* learning methods.

Type	Method	Sketchy (Extended)				TU-Berlin (Extended)			
		mAP @all	Precision @100	Feature Dimension	Retrieval Time (s)	mAP @all	Precision @100	Feature Dimension	Retrieval Time (s)
SBIR	Softmax Baseline	0.099	0.176	4096	3.9×10^{-1}	0.083	0.139	4096	4.7×10^{-1}
	Siamese CNN [49]	0.143	0.183	64	5.2×10^{-3}	0.122	0.153	64	6.3×10^{-3}
	SaN [70]	0.104	0.129	512	4.4×10^{-2}	0.096	0.112	512	5.1×10^{-2}
	GN Triplet [52]	0.211	0.310	1024	8.9×10^{-2}	0.189	0.241	1024	1.4×10^{-1}
	3D Shape [62]	0.062	0.070	64	5.6×10^{-3}	0.057	0.063	64	7.0×10^{-3}
	DSH (64 bits) [38]	0.164	0.227	64 (binary)	6.3×10^{-5}	0.122	0.198	64 (binary)	7.5×10^{-5}
Zero-Shot	CMT [56]	0.084	0.096	300	3.1×10^{-2}	0.065	0.082	300	3.7×10^{-2}
	DeViSE [19]	0.071	0.078	300	3.2×10^{-2}	0.067	0.075	300	3.7×10^{-2}
	SSE [75]	0.108	0.154	100	1.1×10^{-2}	0.096	0.133	220	1.3×10^{-2}
	JLSE [76]	0.126	0.178	100	1.1×10^{-2}	0.107	0.165	220	1.3×10^{-2}
	SAE [31]	0.210	0.302	300	3.1×10^{-2}	0.161	0.210	300	3.7×10^{-2}
	ZSH (64 bits) [67]	0.165	0.217	64 (binary)	6.3×10^{-5}	0.139	0.174	64 (binary)	7.5×10^{-5}
Proposed	**ZSIH (64 bits)**	**0.254**	**0.340**	64 (binary)	6.5×10^{-5}	**0.220**	**0.291**	64 (binary)	7.9×10^{-5}

Table 3. Ablation study. 64-bit mAP@all results of several baselines are shown below.

Description	Sketchy	TU
Kron. layer $\rightarrow$ concatenation	0.228	0.207
Kron. layer $\rightarrow$ MFB [71]	0.236	0.211
Stochastic neuron $\rightarrow$ bit regularization	0.187	0.158
Decoder $\rightarrow$ classifier	0.162	0.133
Without GCNs	0.233	0.171
GCNs $\rightarrow$ word vector fusion	0.219	0.176
$t = 1$ for GCNs	0.062	0.055
$t = 10^{-6}$ for GCNs	0.241	0.202
ZSIH (full model)	**0.254**	**0.220**

art cross-modal hashing works are introduced including CMSSH [4], CMFH [13], SCM [72], CVH [33], SePH [36] and DSH [38], where DSH [38] can also be subjected to an SBIR model and thus is closely related to our work. In addition, CCA [60] is considered as a conventional cross-modal baseline, though it learns real-valued joint representations.

***Zero-Shot* Baselines.** Existing *zero-shot* learning works are not originally designed for cross-modal search. We select a set of state-of-the-art *zero-shot* learning algorithms as benchmarks, including CMT [56], DeViSE [19], SSE [75], JLSE [76], SAE [31] and the *zero-shot* hashing model, *i.e.*, ZSH [67]. For CMT [56], DeViSE [19] and SAE [56], two sets of 300-D embedding functions are trained for sketches as images with the word vectors [44] as the semantic information for nearest neighbour retrieval, and the classifiers used in these works are ignored. SSE [75] and JLSE [76] are based on *seen-unseen* class mapping, so the output embedding sizes are set to 100 and 220 for Sketchy [52] and TU-Berlin [15] dataset respectively. We train two modal-specific encoders of ZSH [67] simultaneously for our task.

Sketch-Image Mapping Baselines. Siamese CNN [49], SaN [70], GN Triplet [52], 3D Shape [62] and DSH [38] are involved as SBIR baselines. We follow the instructions of the original papers to build and train the networks under our *zero-shot* setting. A softmax baseline is additionally introduced, which is based on computing the 4096-D AlexNet [32] feature distances pre-trained on the *seen* classes for nearest neighbour search.

Results and Analysis. The *zero-shot* cross-modal retrieval mean-average precisions (mAP@all) of ZSIH and several hashing baselines are given in Tab. 1, while the corresponding precision-recall (P-R) curves and precision@100 scores are illustrated in Fig. 3. The performance margins between ZSIH and the selected baselines are significant, suggesting the existing cross-modal hashing methods fail to handle our *zero-shot* task. ZSH [67] turns out to be the only well-known *zero-shot* hashing model and it attains relatively better results than other baselines. However, it is originally designed for single-modal data retrieval. DSH [38] leads the SBIR performance under the conventional cross-modal hashing setting, but we observe a dramatic performance drop when extending it to the *unseen* categories. Some retrieval results are provided in Fig. 4. Fig. 5 shows the 32-bit t-SNE [41] results of ZSIH on the training set and test set, where a clearly scattered map on the *unseen* classes can be observed. We also illustrate the retrieval performance w.r.t. the number of *seen* classes in Fig. 5. It can be seen that ZSIH is able to produce acceptable retrieval performance as long as an adequate number of *seen* classes is provided to explore the semantic space.

The comparisons with SBIR and *zero-shot* baselines are shown in Tab. 2, where an akin performance margin to the one of Tab. 1 can be observed. To some extent, the SBIR baselines based on positive-negative samples, *e.g.*, Siamese CNN [49] and GN Triplet [52], have the ability to generalize the learned representations to *unseen* classes. SAE [31] produces closest performance to ZSIH among the *zero-shot* learning baselines. Similar to ZSH [67], these *zero-shot* baselines suffer from the problem of mitigating the modality heterogeneity. Furthermore, most of the methods in Tab. 2 learn real-valued representations, which leads

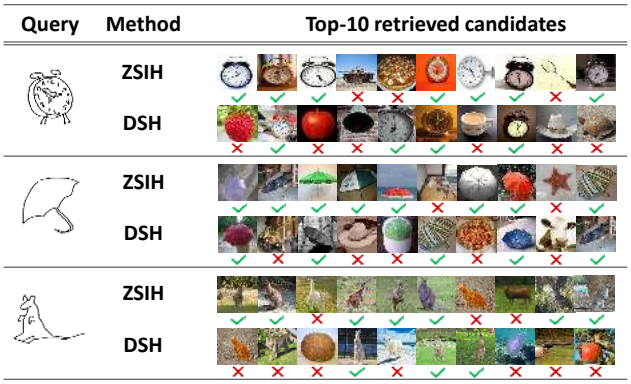

Figure 4. Some top-10 *zero-shot* SBIR results of ZSIH and DSH [38] are shown here according to the hamming distances, where the green ticks indicate correct retrieval candidates and red crosses indicate the wrong ones.

to poor retrieval efficiency when performing nearest neighbour search in the high-dimensional continuous space.

3.4. Ablation study

Some ablation study results are reported in this subsection to justify the plausibility of our proposed model.

Baselines. The baselines in this subsection are built by modifying some parts of the original ZSIH model. To demonstrate the effectiveness of the Kronecker layer for data fusion, we introduce two baselines by replacing the Kronecker layer [22] with the conventional feature concatenation and the multi-modal factorized bilinear pooling (MFB) layer [71]. Regularizing the output bits with quantization error and bit decorrelation loss identical to [37] is also considered as a baseline in replacing the stochastic neurons [9]. The impact of the semantic-aware encoding-decoding design is evaluated by substituting a classifier for the semantic decoder. We introduce another baseline by replacing the graph convolutional layers [30] with conventional fully connected layers. Fusing the word embedding to the multi-modal network is also tested in replacement of graph convolution. Several different hyper-parameter settings of t are also reported.

Results and Analysis. The ablation study results are demonstrated in Tab. 3. We only report the 64-bit mAP on the two datasets for comparison in order to ensure the paper content to be concise. It can be seen that the reported baselines typically underperform the proposed model. Both feature concatenation and MFB [71] produce reasonable retrieval performances, but the figures are still clearly lower than our original design. We speculate this is because the Kronecker layer considerably expands the hidden state dimensionality and therefore, the network is able to store more information for cross-modal hashing. When testing the baseline of bit regularization similar to [37], we experience an unstable training procedure easily leading to overfitting. The quantization error and bit decorrelation loss introduce additional hyper-parameters to the model, making the training procedure hard. Replacing the semantic decoder with a classifier results in a dramatic performance fall as the classifier basically provides no semantic information and fails to generalize knowledge from the *seen* classes to the *unseen* ones. Graph convolutional layer [30] also plays an important role in our model. The mAP drops by about 4% when removing it. Graph convolution enhances hidden representations and knowledge within the neural network by correlating the data points that are semantically close, benefiting our *zero-shot* task. As to the hyper-parameters, a large value of t, *e.g.*, $t = 1$, generally leads to a tightly-related graph adjacency, making data points from different categories hard to be recognized. On the contrary, an extreme small value t, *e.g.*, $t = 10^{-6}$, suggests a sparsely-connected graph with binary-like edges, where only data points from the same category are linked. This is also suboptimal in exploring the semantic relation for *zero-shot* tasks.

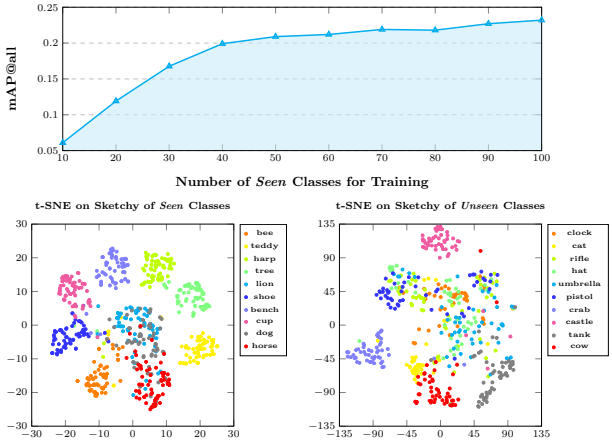

Figure 5. First row: 32-bit ZSIH retrieval performance on Sketchy according to different numbers of *seen* classes used during training. Second row: 32-bit t-SNE [41] scattering results on the Sketchy dataset of the *seen* and *unseen* classes.

4. Conclusion

In this paper, a novel but realistic task of efficient large-scale *zero-shot* SBIR hashing was studied and successfully tackled by the proposed *zero-shot* sketch-image hashing (ZSIH) model. We designed an end-to-end three-network deep architecture to learn shared binary representations and encode sketch/image data. Modality heterogeneity between sketches and images was mitigated by a Kronecker layer with attended features. Semantic knowledge was introduced in assistance of visual information by graph convolutions and a generative hashing scheme. Experiments suggested the proposed ZSIH model significantly outperforms existing methods in our *zero-shot* SBIR hashing task.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 5

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 2

[3] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, 2016. 2

[4] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, 2010. 1, 2, 6, 7

[5] Y. Cao, M. Long, J. Wang, and S. Liu. Collective deep quantization for efficient cross-modal retrieval. In *AAAI*, 2017. 1, 2, 5

[6] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu. Deep visual-semantic hashing for cross-modal retrieval. In *ACM SIGKDD*, 2016. 1, 2

[7] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 2

[8] S. Changpinyo, W.-L. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, 2017. 2

[9] B. Dai, R. Guo, S. Kumar, N. He, and L. Song. Stochastic generative hashing. In *ICML*, 2017. 4, 5, 8

[10] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 2016. 3, 4

[11] B. Demirel, R. G. Cinbis, and N. I. Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *ICCV*, 2017. 2

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[13] G. Ding, Y. Guo, and J. Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, 2014. 1, 2, 6, 7

[14] Z. Ding, M. Shao, and Y. Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR*, 2017. 2

[15] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM Transactions on Graphics*, 31(4):44–1, 2012. 6, 7

[16] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 34(5):482–498, 2010. 1, 2

[17] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics*, 17(11):1624–1636, 2011. 1, 2

[18] V. Erin Liong, J. Lu, Y.-P. Tan, and J. Zhou. Cross-modal deep variational hashing. In *ICCV*, 2017. 2

[19] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2, 3, 7

[20] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2916–2929, 2013. 1

[21] D. K. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011. 4

[22] G. Hu, Y. Hua, Y. Yuan, Z. Zhang, Z. Lu, S. S. Mukherjee, T. M. Hospedales, N. M. Robertson, and Y. Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV*, 2017. 3, 4, 8

[23] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. In *CVIU*, 2013. 1, 2

[24] R. Hu, T. Wang, and J. Collomosse. A bag-of-regions approach to sketch-based image retrieval. In *ICIP*, 2011. 1, 2

[25] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen. Learning discriminative latent attributes for zero-shot classification. In *ICCV*, 2017. 2

[26] Q.-Y. Jiang and W.-J. Li. Deep cross-modal hashing. In *CVPR*, 2017. 1, 2, 5

[27] N. Karessli, Z. Akata, B. Schiele, and A. Bulling. Gaze embeddings for zero-shot image classification. In *CVPR*, 2017. 2

[28] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[29] D. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[30] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3, 4, 8

[31] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017. 2, 3, 7

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 5, 6, 7

[33] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, 2011. 1, 2, 6, 7

[34] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 2

[35] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *CVPR*, 2017. 2

[36] Z. Lin, G. Ding, M. Hu, and J. Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, 2015. 1, 2, 6, 7

[37] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou. Deep hashing for compact binary codes learning. In *CVPR*, 2015. 8

[38] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 1, 2, 5, 6, 7, 8

[39] L. Liu, M. Yu, and L. Shao. Latent structure preserving hashing. *International Journal of Computer Vision*, 122(3):439–457, 2017. 1

[40] Y. Long, L. Liu, Y. Shen, L. Shao, and J. Song. Towards affordable semantic searching: Zero-shot. retrieval via dominant attributes. In *AAAI*, 2018. 2

[41] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 7, 8

[42] D. Mandal, K. N. Chaudhury, and S. Biswas. Generalized semantic preserving hashing for n-label cross-modal retrieval. In *CVPR*, 2017. 2

[43] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber. Multimodal similarity-preserving hashing. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):824–830, 2014. 2

[44] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR Workshop*, 2013. 3, 5, 7

[45] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 4

[46] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 2

[47] K. Pang, Y.-Z. Song, T. Xiang, and T. Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017. 2

[48] S. Parui and A. Mittal. Similarity-invariant sketch-based image retrieval in large databases. In *ECCV*, 2014. 2

[49] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, 2016. 1, 2, 7

[50] J. M. Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *ICIP*, 2014. 1, 2

[51] R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7), 2009. 1

[52] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics*, 35(4):119, 2016. 1, 2, 6, 7

[53] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H. T. Shen. Unsupervised deep hashing with similarity-adaptive and discrete optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2018. 1

[54] Y. Shen, L. Liu, and L. Shao. Unsupervised deep generative hashing. In *British Machine Vision Conference (BMVC)*, 2017. 1

[55] Y. Shen, L. Liu, L. Shao, and J. Song. Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval. In *ICCV*, 2017. 2

[56] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 2, 7

[57] J. Song, Y.-Z. Song, T. Xiang, and T. Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, 2017. 1, 2

[58] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM SIGMOD*, 2013. 2

[59] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017. 1, 2, 3, 5

[60] B. Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005. 6, 7

[61] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966. 4

[62] F. Wang, L. Kang, and Y. Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015. 2, 7

[63] B. Wu, Q. Yang, W.-S. Zheng, Y. Wang, and J. Wang. Quantized correlation hashing for fast cross-modal search. In *IJCAI*, 2015. 2

[64] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 2

[65] X. Xu, F. Shen, Y. Yang, D. Zhang, H. Tao Shen, and J. Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *CVPR*, 2017. 2

[66] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, 2017. 2

[67] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen. Zero-shot hashing via transferring supervised knowledge. In *ACM MM*, 2016. 2, 3, 5, 6, 7

[68] M. Ye and Y. Guo. Zero-shot classification with discriminative semantic representation learning. In *CVPR*, 2017. 2

[69] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 2

[70] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales. Sketch-a-net that beats humans. In *BMVC*, 2015. 1, 2, 7

[71] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, 2017. 4, 7, 8

[72] D. Zhang and W.-J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, 2014. 1, 2, 6, 7

[73] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016. 6

[74] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 2

[75] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. 2, 7

[76] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016. 2, 7

[77] Y. Zhen and D.-Y. Yeung. Co-regularized hashing for multimodal data. In *NIPS*, 2012. 2

[78] R. Zhou, L. Chen, and L. Zhang. Sketch-based image retrieval on a large scale database. In *ACM MM*, 2012. 2