Supervised Hierarchical Cross-Modal Hashing

Changchang Sun
Shandong University
sunchangchang123@gmail.com

Xuemeng Song
Shandong University
sxmustc@gmail.com

Fuli Feng
National University of Singapore
fulifeng93@gmail.com

Wayne Xin Zhao
Renmin University of China
batmanfly@gmail.com

Hao Zhang
Mercari, Inc.
zhtwd@mercari.com

Liqiang Nie
Shandong University
nieliqiang@gmail.com

ABSTRACT

Recently, due to the unprecedented growth of multimedia data, cross-modal hashing has gained increasing attention for the efficient cross-media retrieval. Typically, existing methods on cross-modal hashing treat labels of one instance independently but overlook the correlations among labels. Indeed, in many real-world scenarios, like the online fashion domain, instances (items) are labeled with a set of categories correlated by certain hierarchy. In this paper, we propose a new end-to-end solution for supervised cross-modal hashing, named HiCHNet, which explicitly exploits the hierarchical labels of instances. In particular, by the pre-established label hierarchy, we comprehensively characterize each modality of the instance with a set of layer-wise hash representations. In essence, hash codes are encouraged to not only preserve the layer-wise semantic similarities encoded by the label hierarchy, but also retain the hierarchical discriminative capabilities. Due to the lack of benchmark datasets, apart from adapting the existing dataset FashionVC from fashion domain, we create a dataset from the online fashion platform Ssense consisting of 15,696 image-text pairs labeled by 32 hierarchical categories. Extensive experiments on two real-world datasets demonstrate the superiority of our model over the state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**;

KEYWORDS

Cross-modal Retrieval; Layer-wise Hashing; Hierarchy

ACM Reference Format:
Changchang Sun, Xuemeng Song, Fuli Feng, Wayne Xin Zhao, Hao Zhang, and Liqiang Nie. 2019. Supervised Hierarchical Cross-Modal Hashing. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331229

* Xuemeng Song (sxmustc@gmail.com) and Liqiang Nie (nieliqiang@gmail.com) are corresponding authors.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331229

1 INTRODUCTION

Recent years have witnessed the unprecedented growth of multimedia data on the Internet, thanks to the flourish of multimedia devices (e.g., smart mobile devices) that facilitate people to present one instance with different media types, such as the text and image. Accordingly, it gives rise to the emerging real-world application of cross-media retrieval, which aims to search the semantically similar instances in one modality (e.g., the image) with a query of another modality(e.g., the text). To handle the large-scale multi-modal data efficiently, cross-modal hashing [1, 5, 9, 21–24, 36] has gained increasing attention from researchers due to its remarkable advantages of low time and storage costs. In fact, existing cross-modal hashing methods can be roughly classified into two lines: unsupervised methods [7, 8, 12, 26, 30, 40, 43, 44] and supervised methods [13, 15, 18, 38, 39, 41, 42]. Due to the limitation that the semantic labels of instances cannot be well exploited to strengthen the performance by unsupervised methods, increasing efforts have been dedicated to the supervised manner.

Although existing supervised cross-modal hashing efforts have achieved compelling success [6, 11, 13, 17, 38], they overlooked the semantic correlations among labels of one instance. In fact, in many real-world applications, labels of an instance can be correlated with certain structure. For example, in the online fashion domain, e.g., Ssense[1], to facilitate the user browsing, fashion items are artificially organized within a pre-established category hierarchy and each item is thus labeled with a set of hierarchical categories in different granularity. As shown in Figure 1, item I_1 is annotated by {*Clothing, Skirt, Mini Skirt*}, item I_3 is associated with {*Clothing, Skirt, Long Skirt*}, while item I_7 involves {*Clothing, Jeans, Wide Leg Jeans*}. Apparently, categories at different layers characterize the semantic similarity between fashion items from different perspectives. In terms of the finest-grained layer, items I_1 and I_3 should be semantically dissimilar because of their different specific categories (i.e., "Mini Skirt" and "Long Skirt"), while regarding the less finer-grained layer, I_1 and I_3 can be considered as semantically similar due to their common coarse category of "Skirt". In the light of this, existing studies that treat all categories equally and define the universal inter-modal semantic similarity to supervise the cross-modal hashing can be inappropriate. Beyond that, in this work, we

[1] https://www.ssense.com/.

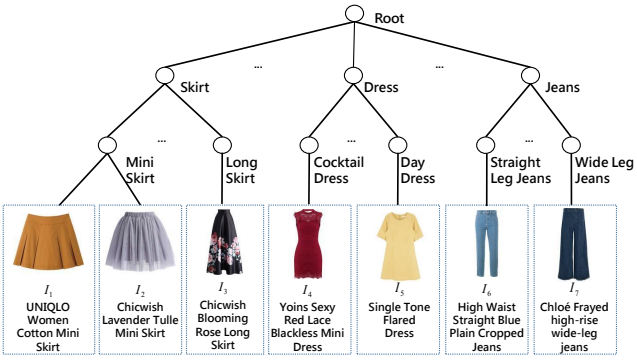


Figure 1: Illustration of the label hierarchy.

aim to boost the performance of supervised hierarchical cross-modal hashing by explicitly exploiting the rich semantic message conveyed by the established category hierarchy [19].

However, fulfilling the task of supervised cross-modal hashing with hierarchical labels is non-trivial due to the following challenges. 1) How to utilize the hierarchical labels to enhance the discriminative power of binary hash codes for the essential semantic encoding constitutes a tough challenge. In a sense, the more discriminative the hash codes are regarding the semantic labels, the more effectively the inter-modal semantic similarity can be measured. 2) How to employ the label hierarchy to guide the cross-modal hashing is another crucial challenge. Undoubtedly, hierarchical labels in different granularity convey more comprehensive semantic information than the traditional independent ones. It is thus inappropriate to resort to the conventional cross-modal hashing that treats all labels equally and measures the semantic similarity among instances simply by counting their common labels. 3) The last challenge lies in the lack of real-world benchmark dataset, whose data points should involve multiple modalities and are hierarchically labeled. Notably, although there are certain hierarchical-labeled datasets, such as the ImageNet [25] and CIFAR [33], they suffer from the limitation of the unimodal data points (e.g., pure images) and thus cannot be adopted for the cross-modal hashing research.

To address the aforementioned challenges, we propose a new supervised hierarchical cross-modal hashing (HiCHNet) method to unify the hierarchical discriminative learning and regularized cross-modal hashing, as shown in Figure 2. In particular, HiCHNet is comprised of an end-to-end dual-path neural network, where each path refers to one modality. To take full advantage of the pre-established label hierarchy, we first characterize each modality of the instance with a set of layer-wise hash representations, corresponding to categories in different granularity. Thereafter, on one hand, we impose the representations of different layers to be discriminative for their corresponding categories. On the other hand, we introduce the layer-wise regularizations as to comprehensively preserve the semantic similarities encoded by the hierarchy. Ultimately, the final binary hash codes, derived from the concatenation of layer-wise hash codes, are encouraged to retain the hierarchical discriminative capabilities and preserve the layer-wise semantic similarities simultaneously. As for the lack of benchmark dataset, we first recognize an existing publicly available dataset FashionVC [31], originally constructed in the context of complementary clothing matching [31], and naturally adapt it for the hierarchical cross-modal hashing. Meanwhile, we further build a benchmark dataset consisting of 15, 696 image-text pairs from the global online fashion platform Ssense, labeled by 32 hierarchical categories. Extensive experiments on two real-world datasets demonstrate the superiority of our model over the state-of-the-art methods.

Our main contributions can be summarized in threefold:

- To the best of our knowledge, this is the first attempt to tackle the real-world problem of cross-modal hashing with hierarchical labels, which has especially great demand in the fashion domain.
- We propose a novel supervised hierarchical cross-modal hashing framework, which is able to seamlessly integrate the hierarchical discriminative learning and the regularized cross-modal hashing.
- We build a large-scale benchmark dataset from the global fashion platform Ssense, which consists of 15, 696 image-text pairs. Extensive experiments demonstrate the superiority of HiCHNet over the state-of-the-art methods. As a byproduct, we have released the datasets, codes, and involved parameters to benefit other researchers[2].

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work and Section 3 details the proposed model. Experimental results and analyses on two datasets are presented in Section 4, followed by our concluding remarks and future work in Section 5.

2 RELATED WORK

Existing cross-modal hashing methods can be roughly divided into two categories: unsupervised and supervised methods.

Unsupervised methods [8, 10, 12, 30, 43] focus on learning hash functions by exploiting the intra- and inter-modality relations with unlabeled training data. For example, Song et al. [30] proposed a novel inter-media hashing (IMH) model to linearly project the heterogeneous data sources into a common Hamming space by co-regularizing the inter- and intra-media consistency. To overcome the limitation of linear projections, Zhou et al. [43] presented the latent semantic sparse hashing (LSSH) model, where the high-level latent semantic information conveyed by the images and texts is well-captured by employing Sparse Coding and Matrix Factorization. Noting that the quantization errors should be punished to improve the performance, Irie et al. [12] proposed the alternating co-quantization (ACQ) scheme that alternately seeks the binary quantizers for each modality by jointly solving the subspace learning and binary quantization. Even integrated with the simple CCA [10], ACQ can boost the retrieval performance significantly. Overall, although existing unsupervised methods have achieved promising performance, they neglected the value of the existing semantic label information and hence suffer from the inferior performance.

Supervised methods [13, 15, 15, 17, 28, 38, 39, 41] work on leveraging the semantic labels of training data as the supervision to guide the hash codes learning and boost the performance. For example, Zhang et al. [39] put forward an effective semantic

[2]https://drive.google.com/drive/folders/1v1qu3AwSPucmjfuw2r81ORD9HjyK8Hdr.

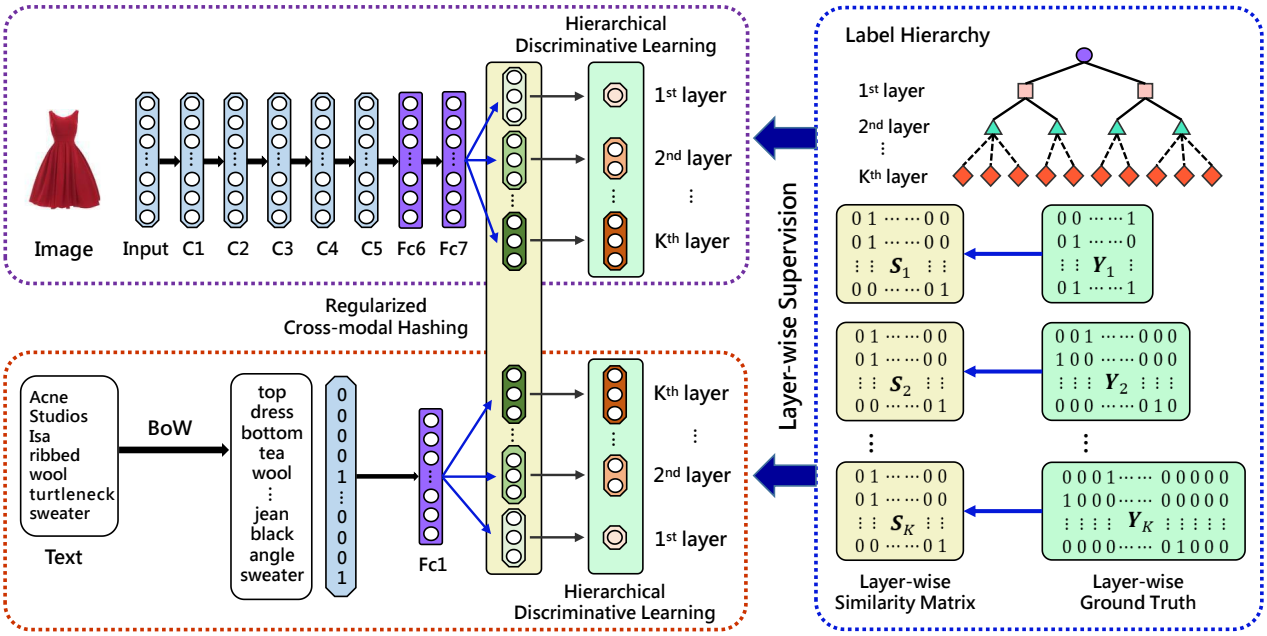


Figure 2: Illustration of the proposed scheme. HiCHNet characterizes each modality of the instance with a set of layer-wise hash representations via the corresponding neural network, which is regularized to retain the hierarchical discriminative capability and hence preserve the layer-wise semantic similarities derived from the ground truth labels.

correlation maximization (SCM) method to seamlessly integrate the semantic labels into the hashing learning. In addition, to capture the underlying semantic information, Yu et al. [38] introduced a two-stage discriminative coupled dictionary hashing (DCDH) model to jointly learn the coupled dictionaries and hash functions for both modalities. Furthermore, arguing that the semantic affinities can be used to guide the hashing, Lin et al. [17] formulated a semantics-preserving hashing (SePH) paradigm where the probability distribution generated from semantic affinities is approximated via minimizing the Kullback-Leibler divergence. It is worth noting that the above methods mainly rely on the hand-crafted features, which inevitably leads to the separate feature extraction and hash codes learning procedures. To overcome this drawback, Jiang et al.[13] established an end-to-end deep cross-modal hashing (DCMH) framework with deep neural networks, one for each modality to perform feature learning from scratch. In spite of the compelling success achieved by these methods in general cases, far too little attention has been paid to the real-world domains with hierarchical labels like the fashion domain. In fact, it is inappropriate to directly apply existing supervised methods that treat all labels equally and overlook the hierarchical relatedness among them.

In fact, the concept of hierarchy has been noticed by many researchers [16, 27, 32, 35]. For example, Song et al. [32] explored the hierarchical relatedness among user interests and proposed a structure-constrained multi-source multi-task learning scheme for the user interest inference. For the hashing domain, Wang et al. [35] presented a supervised hierarchical deep hashing method in the context of unimodal hashing. Nevertheless, the potential of hierarchical labels in cross-modal hashing has not been well validated, which is the major concern of this paper.

3 PRELIMINARIES

We first introduce the necessary notations throughout the paper, and then define the studied task.

3.1 Notation

Suppose that we have a set of N instances $\mathcal{E} = \{e_i\}_{i=1}^{N}$ labeled by a set of categories that are not independent but correlated with a hierarchy of $(K+1)$ layers. We compile the $(K+1)$ layers from top to bottom with the index set $\{0, 1, \cdots, K\}$, where the 0-th layer corresponds to the root node. Let c^k denote the number of nodes at the k-th layer. As for the i-th instance $e_i = (\mathbf{v}_i, \mathbf{t}_i, \mathcal{Y}_i)$, $\mathbf{v}_i \in \mathbb{R}^{d_v}$ and $\mathbf{t}_i \in \mathbb{R}^{d_t}$ stand for the original image and text feature vectors, where d_v and d_t represent the respective feature dimensions. $\mathcal{Y}_i = \left\{\mathbf{y}_i^k\right\}_{k=1}^{K}$ denotes the set of label vectors for e_i, where $\mathbf{y}_i^k = [y_{i1}^k, y_{i2}^k, \cdots, y_{ic_k}^k]^T \in \{0,1\}^{c^k}$ is the label vector pertaining to the categories of the k-th layer[3]. In particular, $y_{ij}^k = 1$, if the i-th instance e_i is labeled with the j-th category at the k-th layer, otherwise $y_{ij}^k = 0$. For simplicity, we define $\mathbf{Y}^k = [\mathbf{y}_1^k, \mathbf{y}_2^k, \cdots, \mathbf{y}_N^k] \in \{0,1\}^{c^k \times N}$ as the label matrix for the k-th layer of all instances in $\mathcal{E}$. Moreover, according to the label hierarchy, we introduce a set of K layer-wise inter-modal similarity matrices $\mathcal{S} = \left\{\mathbf{S}^k\right\}_{k=1}^{K}$, where $\mathbf{S}^k \in \{0,1\}^{N \times N}$ corresponds to the similarities among all instances regarding categories in the k-th layer. In particular, the (i,j)-th entry $S_{ij}^k = 1$ if the image of instance e_i and text of instance e_j share the identical label for the k-th layer (i.e., $\mathbf{y}_i^k = \mathbf{y}_j^k$), otherwise $S_{ij}^k = 0$. Table 1 summarizes the main notations used in this paper.

[3]Here, we do not consider the 0-th layer of the root node.

Table 1: Summary of the main notations.

Notation	Explanation
K	Num. of layers in the hierarchy except the root.
L	The length of the hash codes.
e_i	The i-th instance.
$\mathbf{y}_i^k$	Label vector of e_i pertaining to the k-th layer.
$\mathbf{S}^k$	Inter-modal similarity matrix of the k-th layer.
f^v (f^t)	Hash function for the image (text) modality.
$\mathbf{\Theta}_v$ ($\mathbf{\Theta}_t$)	Parameters of f^v (f^t).
$\mathbf{v}_i$ ($\mathbf{t}_i$)	Original image (text) feature vector of e_i.
$\mathbf{b}_{v_i}$ ($\mathbf{b}_{t_i}$)	Image (text) hash codes of e_i.
$\mathbf{h}_{v_i}$ ($\mathbf{h}_{t_i}$)	Image (text) hash representation of e_i.

3.2 Problem Formulation

In this work, we aim to devise an end-to-end supervised hierarchical cross-modal hashing learning scheme to obtain the accurate image and text L-bit hash codes for the i-th instance, namely, $\mathbf{b}_{v_i} \in \{-1,1\}^L$ and $\mathbf{b}_{t_i} \in \{-1,1\}^L$. Based on the hash codes, we can measure the inter-modal similarities using the Hamming distance as $dis_H(\mathbf{b}_{v_i}, \mathbf{b}_{t_j}) = \frac{1}{2}(L - \mathbf{b}_{v_i}^T \mathbf{b}_{t_j})$ and hence perform the cross-modal retrieval.

To simplify the presentation, we focus on the cross-modal retrieval for the bimodal data (i.e., the image and text). Without losing the generality, our task can be easily extended to the scenarios with multiple modalities. In particular, we aim to learn hash codes for the image and text modalities (i.e., $\mathbf{b}_{v_i} = sgn(f^v(\mathbf{v}_i; \mathbf{\Theta}_v))$ and $\mathbf{b}_{t_i} = sgn(f^t(\mathbf{t}_i; \mathbf{\Theta}_t))$), respectively. $sgn(\cdot)$ is the element-wise sign function, which outputs " + 1" for positive real numbers and " − 1" for negative ones. Here, f^v and f^t refer to the hashing networks with parameters $\mathbf{\Theta}_v$ and $\mathbf{\Theta}_t$ to be learned.

4 THE PROPOSED MODEL

In this section, we present the proposed HiCHNet, as the major novelty, which is able to effectively leverage the label hierarchy information for improving the learning of cross-modal hash codes. In particular, we first set up *layer-wise hash representations* for capturing semantic characteristics in different granularity and then enhance their discriminative power with *hierarchical discriminative learning*, and finally instruct the hashing learning with *regularized cross-modal hashing*.

4.1 Layer-wise Hash Representation

Intuitively, as different modalities of one instance are semantically correlated, an effective hashing model should be able to preserve the similarity between different modalities for the same instance. Nevertheless, it is inadvisable to directly measure the inter-modal similarity from the original heterogeneous feature spaces.

Inspired by the huge success of the representation learning, we adopt deep neural networks to obtain more powerful image and text representations. Regarding the image modality, we utilize the convolution neural network (CNN) adapted from [4] consisting of five convolution layers followed by two fully-connected layers. In particular, given the i-th instance, we feed its original image feature $\mathbf{v}_i$ (i.e., the pixel vector) to the CNNs, and adopt the fc7 layer output as the image representation $\tilde{\mathbf{v}}_i$. As for the text modality, in the similar manner, we employ a neural network comprising one fully-connected layer [20] to transform the original text feature vector $\mathbf{t}_i$ into the text representation $\tilde{\mathbf{t}}_i$.

Having obtained the image and text representations of instances, we can perform the respective projection from the representation space to the Hamming space and derive the hash codes for each modality. To fully exploit the hierarchy, our idea is to set layer-wise representations for each modality corresponding to the category layers of the hierarchy with different granularity. Formally, we equally divide the general L-bit hash codes into K layer-wise hash codes, namely, $\mathbf{b}_{v_i} = [\mathbf{b}_{v_i}^1, \mathbf{b}_{v_i}^2, \cdots, \mathbf{b}_{v_i}^K]$ and $\mathbf{b}_{t_i} = [\mathbf{b}_{t_i}^1, \mathbf{b}_{t_i}^2, \cdots, \mathbf{b}_{t_i}^K]$, where $\mathbf{b}_{v_i}^k$ and $\mathbf{b}_{t_i}^k$ refer to the image and text hash codes of instance e_i regarding the k-th layer.

For the image modality, we feed the image representation $\tilde{\mathbf{v}}_i$ to K separate networks simultaneously, each of which comprises one fully-connected layer as follows,

$$\mathbf{h}_{v_i}^k = s(\mathbf{W}_v^k \tilde{\mathbf{v}}_i + \mathbf{g}_v^k), \quad k = 1, \cdots, K, \tag{1}$$

where $\mathbf{h}_{v_i}^k \in \mathbb{R}^{z_k}$ refers to the image hash representation for the k-th layer with the dimension of z_k, and $\mathbf{W}_v^k$ and $\mathbf{g}_v^k$ are the weight matrix and bias vector, respectively. And $s : \mathbb{R} \mapsto \mathbb{R}$ is a non-linear function applied element wise[4]. Then, based on the set of image hash representations for the i-th instance $\left\{\mathbf{h}_{v_i}^k\right\}_{k=1}^K$, we can get the binary layer-wise image hash codes as follows,

$$\mathbf{b}_{v_i}^k = sgn(\mathbf{h}_{v_i}^k), \quad k = 1, \cdots, K, \tag{2}$$

where $\mathbf{b}_{v_i}^k \in \{-1,1\}^{z_k}$. In a similar manner, we can derive the layer-wise text hash representations $\left\{\mathbf{h}_{t_i}^k\right\}_{k=1}^K$ and binary text hash codes $\left\{\mathbf{b}_{t_i}^k\right\}_{k=1}^K$ for the i-th instance.

4.2 Hierarchical Discriminative Learning

In a sense, as to comprehensively encode the necessary semantic information from the hierarchy, the layer-wise hash codes, which can be regarded as the projected representations for instances in the Hamming space, should be discriminative towards the semantic classification in different granularity over the hierarchy. Towards this end, we introduce K layer-wise multiple classification tasks simultaneously. For the k-th multi-classification, we particularly take the k-th layer hash representations as the input and labels regarding the k-th layer of the hierarchy as the ground truth. For simplicity, we take the discriminative learning of the image modality as an example and that of the text modality can be effortlessly achieved in the same manner.

In particular, we feed K layer-wise image hash representations of the i-th instance to K multi-layer perceptrons as follows,

$$\mathbf{p}_{v_i}^k = softmax(\mathbf{U}_v^k \mathbf{h}_{v_i}^k + \mathbf{q}_v^k), \quad k = 1, \cdots, K, \tag{3}$$

where $\mathbf{p}_{v_i}^k \in \mathbb{R}^{c_k}$ refers to the output class distribution pertaining to the k-th layer of the hierarchy, $\mathbf{U}_v^k$ and $\mathbf{q}_v^k$ are the weight matrix and bias vector, respectively. Considering that categories in different granularity may contribute differently to the discriminative regularization, we incorporate the layer confidence for each layer.

[4] In this work, we use the hyperbolic tangent function.

Ultimately, adopting the negative log-likelihood loss for the K layer-wise discriminative classifications, we have,

$$\Psi_h = -\sum_{k=1}^{K}\rho_k\sum_{i=1}^{N}\left[(\mathbf{y}_i^k)^T\log(\mathbf{p}_{v_i}^k)+(\mathbf{y}_i^k)^T\log(\mathbf{p}_{t_i}^k)\right], \tag{4}$$

where ρ_k refers to the confidence of the k-th layer and $log(\cdot)$ is the element-wise logarithm function.

4.3 Regularized Cross-modal Hashing

Above, we have considered the layer-wise correspondence between the hash codes and the category hierarchy. In this part, we first employ the layer-wise regularizations for comprehensively preserving the semantic similarities between different modalities. Then, we incorporate the binarization difference penalizing to further enhance the cross-modal hashing learning.

Semantic Similarity Preserving. To ensure the performance of cross-modal hashing, one major concern is to preserve the inter-modal semantic similarity between two instances when they are mapped from the original representation space to the Hamming space. Consequently, it is desirable to maximize the Hamming distance between two instances whose semantic similarity is 0, while minimizing that with the similarity of 1. Traditionally, existing researches treat all categories independently and only define the universal semantic similarity to preserve, where the category hierarchy has not been utilized. In fact, the hash codes of instances from correlated categories, (e.g., "Long Skirt" and "Mini Skirt" are correlated by sharing the same ancestor category "Skirt") tend to be more similar than that from uncorrelated ones (e.g., "Wide Leg Jeans" and "Mini Skirt"). Accordingly, we also define the semantic similarity in the layer-wise manner as follows,

$$\phi_{ij}^k = \frac{1}{2}(\mathbf{h}_{v_i}^k)^T\mathbf{h}_{t_j}^k, \tag{5}$$

where ϕ_{ij}^k denotes the semantic similarity between image of instance $\mathbf{e}_i$ and text of instance $\mathbf{e}_j$ regarding the k-th layer. The hash representations $\mathbf{h}_{v_i}^k$ and $\mathbf{h}_{t_j}^k$ can be treated as the continuous surrogates of the binary hash codes $\mathbf{b}_{v_i}^k$ and $\mathbf{b}_{t_j}^k$, $k = 1, 2, \cdots, K$, $i, j = 1, 2, \cdots, N$, respectively.

Similar to [13], we encourage ϕ_{ij}^k to approximate the binary ground truth S_{ij}^k as follows,

$$L(\phi_{ij}^k|S_{ij}^k) = \sigma(\phi_{ij}^k)^{S_{ij}^k}(1-\sigma(\phi_{ij}^k))^{(1-S_{ij}^k)}, \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function. Besides, considering that labels in different granularity at different layers may possess different capabilities regarding the semantic similarity regularization, we further introduce the layer confidence. Simple algebra computations enable us to reach the following objective function,

$$\Gamma_1 = -\sum_{k=1}^{K}\tau_k\sum_{i,j=1}^{N}(S_{ij}^k\phi_{ij}^k-\log(1+e^{\phi_{ij}^k})), \tag{7}$$

where τ_k denotes the layer confidence for the k-th layer.

Binarization Difference Penalizing. Apart from the semantic preserving regularization on $\mathbf{h}_{v_i}^k$'s and $\mathbf{h}_{t_j}^k$'s, we further regularize the binarization differences between $\mathbf{h}_{v_i}^k$ and $\mathbf{b}_{v_i}^k$, $\mathbf{h}_{t_i}^k$ and $\mathbf{b}_{t_i}^k$, respectively, as to derive the optimal continuous surrogates of the binary hash codes. For simplicity, we introduce two sets of layer-wise hash representation matrices $\left\{\mathbf{H}_v^k\right\}_{k=1}^K$ and $\left\{\mathbf{H}_t^k\right\}_{k=1}^K$ for the image and text modalities, respectively, where $\mathbf{H}_v^k = [\mathbf{h}_{v_1}^k, \mathbf{h}_{v_2}^k, \cdots, \mathbf{h}_{v_N}^k] \in \mathbb{R}^{z_k\times N}$ and $\mathbf{H}_t^k = [\mathbf{h}_{t_1}^k, \mathbf{h}_{t_2}^k, \cdots, \mathbf{h}_{t_N}^k] \in \mathbb{R}^{z_k\times N}$. Moreover, we can also define two sets of binary layer-wise hash code matrices $\mathcal{B}_v = \left\{\mathbf{B}_v^k\right\}_{k=1}^K$ and $\mathcal{B}_t = \left\{\mathbf{B}_t^k\right\}_{k=1}^K$, where $\mathbf{B}_v^k = [\mathbf{b}_{v_1}^k, \mathbf{b}_{v_2}^k, \cdots, \mathbf{b}_{v_N}^k] \in \{-1,1\}^{z_k\times N}$ and $\mathbf{B}_t^k = [\mathbf{b}_{t_1}^k, \mathbf{b}_{t_2}^k, \cdots, \mathbf{b}_{t_N}^k] \in \{-1,1\}^{z_k\times N}$. The binarization difference regularization thus can be written as follows,

$$\Gamma_2 = \sum_{k=1}^{K}\left(\left\|\mathbf{B}_v^k-\mathbf{H}_v^k\right\|_F^2+\left\|\mathbf{B}_t^k-\mathbf{H}_t^k\right\|_F^2\right), \tag{8}$$

where $\left\|\cdot\right\|_F$ denotes the Frobenius norm.

Consequently, we have the following objective function towards the hierarchical cross-modal hashing,

$$\begin{aligned}\Psi_r = &-\sum_{k=1}^{K}\tau_k\sum_{i,j=1}^{N}\left(S_{ij}^k\phi_{ij}^k-\log(1+e^{\phi_{ij}^k})\right)\\ &+\alpha\left(\left\|\mathbf{B}_v^k-\mathbf{H}_v^k\right\|_F^2+\left\|\mathbf{B}_t^k-\mathbf{H}_t^k\right\|_F^2\right)\\ &+\beta\left(\left\|\mathbf{H}_v^k\mathbf{a}\right\|_2^2+\left\|\mathbf{H}_t^k\mathbf{a}\right\|_2^2\right),\end{aligned} \tag{9}$$

where α and β are the nonnegative tradeoff parameters and $\mathbf{a} = [1, 1, \cdots, 1]^T \in \mathbb{R}^N$, and $\left\|\cdot\right\|_2$ denotes the Euclidean norm. The last term is to balance the learned hash codes and maximize the information conveyed by each bit of the codes [13].

Notably, to bridge the semantic gap between different modalities more effectively and boost the performance of the cross-modal hashing, we adopt the unified binary hash codes (i.e., $\mathbf{B}_v^k = \mathbf{B}_t^k = \mathbf{B}^k$) in the training procedure. Towards this end, we slightly adapt the

Algorithm 1 Supervised Hierarchical Cross-Modal Hashing

Input: Instance set $\mathcal{E}$, similarity matrix set $\mathcal{S}$.

Output: Parameters $\mathbf{\Theta}_v$ and $\mathbf{\Theta}_t$, hash code matrices $\left\{\mathbf{B}^k\right\}_{k=1}^K$.

Initialization

Initialize parameters: α, β, γ, τ_k, ρ_k, $\mathbf{\Theta}_v$, $\mathbf{\Theta}_t$, mini-batch size: m, iteration number: $M = \lceil N/m \rceil$.

repeat

  for $iter = 1, 2, \cdots, M$ **do**

    Randomly sample a batch of m instances from $\mathcal{E}$.

    Feed them into f^v and compute $\left\{\mathbf{H}_v^k\right\}_{k=1}^K$.

    Update $\mathbf{\Theta}_v$ according to Eqn. (11) and (12).

  end for

  for $iter = 1, 2, \cdots, M$ **do**

    Randomly sample a batch of m instances from $\mathcal{E}$.

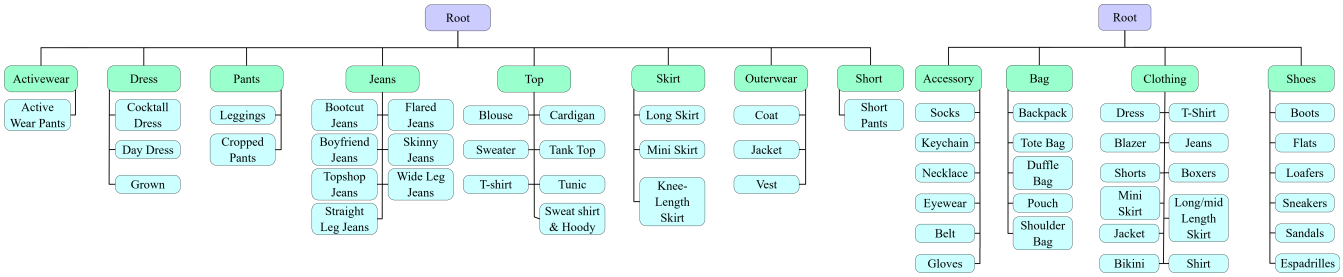
    Feed them into f^t and compute $\left\{\mathbf{H}_t^k\right\}_{k=1}^K$.

    Update $\mathbf{\Theta}_t$ according to Eqn. (11) and (12).

  end for

  Compute $\left\{\mathbf{B}^k\right\}_{k=1}^K$ according to Eqn. (13).

until Convergence



(a) FashionVC. (b) Ssense.

Figure 3: Label hierarchy of datasets FashionVC and Ssense.

derivation of the binary hash code matrix $\mathbf{B}^k$ as follows,

$$\mathbf{B}^k = sgn\left(\mathbf{H}_v^k + \mathbf{H}_t^k\right). \tag{10}$$

4.4 Joint Model and Optimization

Integrating the two key components of the hierarchical discriminative learning and regularized cross-modal hashing, we reach the final objective formulation Ψ as follows,

$$\min_{\mathbf{B}^k,\mathbf{\Theta}_v,\mathbf{\Theta}_t} \gamma\Psi_h + (1-\gamma)\Psi_r, \tag{11}$$

where γ is the nonnegative tradeoff parameter. Overall, we expect the layer-wise hash codes to be discriminative for the hierarchical semantic classification as well as effective towards the cross-modal hashing. It is worth noting that although we assume that both modalities of each instance are observed in the training phase, our scheme can also be easily extended to handle other scenarios, where some training instances miss certain modality. Moreover, once the model has been trained, we can directly use f^v and f^t to generate hash codes for any instance with either one or two modalities and fulfill the cross-modal retrieval task.

We adopt the alternating optimization strategy to solve $\mathbf{B}^k$, $\mathbf{\Theta}_v$ and $\mathbf{\Theta}_t$, where we optimize one variable while fixing the other two in each iteration and keep the iterative procedure until the objective function converges. Due to that $\mathbf{\Theta}_v$ and $\mathbf{\Theta}_t$ share the similar optimization, here we take $\mathbf{\Theta}_v$ as an example. We first calculate the derivative of Ψ with respect to $\mathbf{h}_{v_i}^k$ as $\frac{\partial\Psi}{\partial\mathbf{h}_{v_i}^k} =$

$$\frac{1}{2}\sum_{j=1}^{N}(\sigma(\phi_{ij}^k)\mathbf{h}_{t_j}^k - S_{ij}^k\mathbf{h}_{t_j}^k) + 2\alpha(\mathbf{h}_{v_i}^k - \mathbf{b}_{v_i}^k) + 2\beta\mathbf{H}_v^k\mathbf{a}, \tag{12}$$

where $k = 1,\cdots,K$, and $\frac{\partial\Psi}{\partial\mathbf{\Theta}_v}$ can be derived from $\frac{\partial\Psi}{\partial\mathbf{h}_{v_i}^k}$ using the chain rule. As for the binary hash code matrix $\mathbf{B}^k$, we have

$$\frac{\partial\Psi}{\partial\mathbf{B}^k} = 2\alpha(2\mathbf{B}^k - \mathbf{H}_v^k - \mathbf{H}_t^k), \tag{13}$$

where $k = 1,\cdots,K$. Indeed, $\frac{\partial\Psi}{\partial\mathbf{h}_{v_i}^k}$, $\frac{\partial\Psi}{\partial\mathbf{h}_{t_j}^k}$ and $\frac{\partial\Psi}{\partial\mathbf{B}^k}$ enable us to solve all the parameters via the stochastic gradient descent (SGD) with back-propagation. The overall procedure of the alternating optimization is briefly summarized in Algorithm 1. As each iteration can decrease Ψ, whose lower bound is zero, we can guarantee the convergence of Algorithm 1 [13, 15, 34].

Table 2: Statistics of our datasets.

	FashionVC	Ssense
Training Set	16,862	13,696
Retrieval Set	16,862	13,696
Query Set	3,000	2,000
Total Labels	35	32
The First Layer Labels	8	4
The Second Layer Labels	27	28

5 EXPERIMENT

To evaluate the proposed method, we conducted extensive experiments on two real-world datasets by answering the following research questions:

- Does the proposed HiCHNet outperform the state-of-the-art methods?
- What is the component level contribution of HiCHNet?
- What is the effect of the label hierarchy?

In this section, we first introduce the datasets as well as the experimental settings, and then provide the experimental results with detailed discussions over each above research question.

5.1 Datasets

For the evaluation, we utilized two datasets: FashionVC and Ssense, where the former is adapted from an existing dataset and the latter is created by our own.

FashionVC. On one hand, we adopted the public dataset FashionVC [31] originally collected from the online fashion community Polyvore[5] in the context of clothing matching. FashionVC consists of 20,726 multi-modal fashion items (e.g., tops and bottoms), where each fashion item is composed of a visual image with a clean background, a textual description and hierarchical categories in different granularity. The multi-modal and hierarchically-labeled features of FashionVC naturally propel us to adapt it for the purpose of hierarchical cross-modal hashing. Notably, to guarantee the quality of dataset, we manually removed the noisy items with inconsistent labels and filtered out the categories with less than 25 items. Moreover, we noticed that to

[5]Polyvore has been acquired by the global fashion platform Ssense in 2018.

Table 3: The MAP scores of different methods on FashionVC and Ssense, where two retrieval tasks and different hash code lengths are adopted. The last row refers to the performance improvement of HiCHNet over the best baseline. The shallow learning baselines use the SIFT features and the best accuracy is shown in boldface.

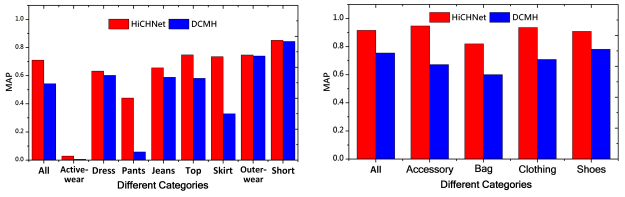
Method	FashionVC								Ssense							
	Image→Text				Text→Image				Image→Text				Text→Image			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CCA	0.150	0.130	0.114	0.103	0.141	0.126	0.112	0.102	0.349	0.292	0.240	0.195	0.355	0.302	0.245	0.195
SCM-Or	0.176	0.128	0.109	0.095	0.159	0.121	0.102	0.091	0.299	0.222	0.169	0.145	0.268	0.181	0.126	0.101
SCM-Se	0.303	0.328	0.355	0.217	0.254	0.288	0.308	0.175	0.489	0.504	0.516	0.498	0.433	0.457	0.476	0.457
DCH	0.224	0.246	0.266	0.312	0.209	0.254	0.300	0.299	0.376	0.419	0.480	0.439	0.434	0.457	0.595	0.611
CDQ	0.456	0.583	0.559	0.621	0.445	0.593	0.538	0.598	**0.769**	0.778	0.856	0.840	**0.766**	0.802	0.864	0.806
SSAH	0.609	0.661	0.703	0.391	0.724	0.794	0.814	0.433	0.443	0.374	0.220	0.149	0.449	0.365	0.223	0.113
DCMH	0.502	0.579	0.603	0.623	0.589	0.638	0.665	0.679	0.608	0.635	0.664	0.690	0.606	0.630	0.666	0.687
HiCHNet	**0.611**	**0.688**	**0.721**	**0.715**	**0.818**	**0.871**	**0.884**	**0.885**	0.696	**0.822**	**0.880**	**0.895**	0.676	**0.838**	**0.873**	**0.916**
↑	0.2%	2.7%	1.8%	9.2%	9.4%	7.7%	7.0%	20.6%	–	4.4%	2.4%	5.5%	–	3.6%	0.9%	11.0%

Table 4: The MAP scores of different methods on two datasets, where the VGG-F feature is used for shallow learning baselines. The best accuracy is shown in boldface.

Method	FashionVC								Ssense							
	Image→Text				Text→Image				Image→Text				Text→Image			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CCA	0.217	0.197	0.182	0.162	0.243	0.224	0.208	0.186	0.460	0.494	0.390	0.301	0.521	0.567	0.472	0.377
SCM-Or	0.256	0.176	0.141	0.121	0.278	0.185	0.133	0.110	0.421	0.295	0.226	0.184	0.372	0.255	0.167	0.123
SCM-Se	0.429	0.462	0.373	0.486	0.522	0.564	0.431	0.593	0.538	0.577	0.584	0.574	0.557	0.600	0.606	0.599
DCH	0.356	0.525	0.566	0.602	0.420	0.586	0.627	0.705	0.600	0.664	0.756	0.589	0.670	0.779	0.843	0.734
CDQ	0.456	0.583	0.559	0.621	0.445	0.593	0.538	0.598	**0.769**	0.778	0.856	0.840	**0.766**	0.802	0.864	0.806
SSAH	0.609	0.661	0.703	0.391	0.724	0.794	0.814	0.433	0.443	0.374	0.220	0.149	0.449	0.365	0.223	0.113
DCMH	0.502	0.579	0.603	0.623	0.589	0.638	0.665	0.679	0.608	0.635	0.664	0.690	0.606	0.630	0.666	0.687
HiCHNet	**0.611**	**0.688**	**0.721**	**0.715**	**0.818**	**0.871**	**0.884**	**0.885**	0.696	**0.822**	**0.880**	**0.895**	0.676	**0.838**	**0.873**	**0.916**
↑	0.2%	2.7%	1.8%	9.2%	9.4%	7.7%	7.0%	18.0%	–	4.4%	2.4%	5.5%	–	3.6%	0.9%	11.0%

facilitate users to explore the website, Polyvore separates all the fashion items into two categories: "men's fashion" and "women's fashion". Nevertheless, for the same category (e.g., the 'T-shirt' or 'Jeans'), the men's and women's garments can be highly visually similar and difficult to be distinguished. We thus excluded these two general categories, and merged the common sub-categories accordingly. Finally, we obtained 19,862 hierarchically-labeled multi-modal training instances, where the label hierarchy consists of 35 categories with two layers, as shown in Figure 3(a).

Ssense. On the other hand, we created our own dataset by crawling the global online fashion platform Ssense, where similar to Polyvore, fashion items with rich multi-modal data are annotated by a set of pre-defined hierarchical categories. In particular, we collected all the fashion items on Ssense, including their visual images, textural descriptions and hierarchical labels during the period of December 14 to 16, 2018 and obtained 25,947 raw labeled image-text instances. Pertaining to the dataset preprocessing, we removed the noisy instances whose images involve multiple items (e.g., both a coat and a dress appear in one image). Due to the similar concern with FashionVC, we discarded the categories 'men' as well as 'women'. Thereafter, we filtered out the categories with less than 70 instances to avoid the unbalanced dataset. Ultimately, we obtained the benchmark dataset of 15,696 image-text instances labeled with a two-layer category hierarchy, as shown in Figure 3(b).



(a) FashionVC (b) Ssense

Figure 4: Performance of HiCHNet and DCMH on different categories of FashionVC and Ssense in the task of "Text→Image".

5.2 Experimental Settings

Evaluation. In this work, we evaluated our method in the context of two classic cross-modal retrieval tasks: querying the image database with textual descriptions ("Text→Image") and querying the text database with given image examples ("Image→Text"). Towards this end, for FashionVC, we randomly sampled 15% of the dataset to form the query set and kept the remaining 85% as the training and retrieval set. Similarly, regarding Ssense, we took

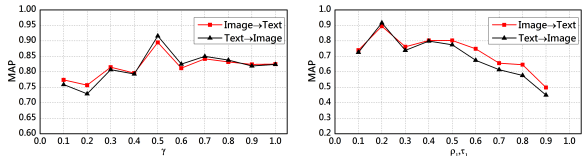
13% of the dataset as the query database and the rest as the training and retrieval database. Statistics regarding our datasets are listed in Table 2. For each cross-modal retrieval task, we utilized the conventional retrieval protocol of Hamming ranking, where the mean average precision (MAP) [37] was employed to measure the performance.

Baselines. We compared the proposed HiCHNet with six state-of-the-art baselines including five supervised methods: SCM [39], CDQ [2], DCH [37], SSAH [15] and DCMH [13], and one unsupervised method: CCA [10]. As SCM has both orthogonal projection and sequential learning modes, we accordingly derived two methods: SCM-Or and SCM-Se. Moreover, we utilized the finest-grained categories as the ground truth for the supervised methods. Notably, CDQ, SSAH, DCMH and HiCHNet are deep learning-based methods, while the others are the shallow learning methods. We thus directly fed the raw images and texts as input for CDQ, SSAH, DCMH and HiCHNet, where we unified the image size to $224 \times 224 \times 3$ by proportionally resizing and padding. The raw text of each instance in FashionVC and Ssense was respectively represented as a 2685-D and 4945-D bag-of-words vector. Regarding shallow learning methods, for fairness, we used both the hand-crafted 500-D SIFT [14] feature and the deep VGG-F feature [3] extracted from the neural networks pre-trained on the ImageNet. Ultimately, we implemented HiCHNet with the open source deep learning software library Tensorflow, and all baselines using the implementations provided by the original authors.

Parameter Setting. We initialized the first seven layers of the deep neural networks for image representation ($\tilde{\mathbf{v}}_i$'s) extraction with the pre-trained VGG-F network parameters [4], while all the other parameters in the neural networks were initialized randomly. Pertaining to the optimization, we utilized the stochastic gradient descent (SGD) [29] with the momentum factor as 0.9. The grid search strategy was adopted to determine the optimal values for the regularization parameters (i.e., τ_k's, ρ_k's, α, β and γ). In addition, we empirically set the batch-size to be 128 and the maximum number of iterations as 500 to ensure the convergence.

5.3 On Model Comparison (RQ1)

To comprehensively evaluate the proposed HiCHNet, we first reported the MAP results of different methods in Tables 3 and 4, where the shallow learning baselines are based on the SIFT feature and deep VGG-F feature, respectively. From Tables 3 and 4, we can draw the following observations: 1) our HiCHNet consistently outperforms all the other baselines with different hash code lengths on FashionVC. In particular, with the best baseline, HiCHNet achieves the significant average improvement of 3.5%, 10.9%, 4.1% and 5.2% in both tasks of "Image→Text" and "Text→Image" on FashionVC and Ssense, respectively. This can be attributed to the fact that HiCHNet is able to not only retain the discriminative capability of the hash codes but also preserve more accurate and comprehensive layer-wise semantic similarities between different modalities. 2) Overall, the performance of HiCHNet is significantly better than all baselines, except for the CDQ on Ssense with the hash code lengths of 16. 3) Meanwhile, the shallow learning baselines with VGG-F feature surpass those based on the SIFT feature, which shows the advantage of the deep learning in feature extraction. 4)



(a) γ (b) ρ_1, τ_1

Figure 5: Sensitivity analysis of the hyper-parameters.

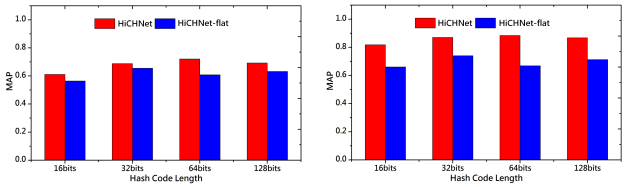
Interestingly, nearly all methods show the better performance on Ssense than FashionVC. One possible explanation is that instances in Ssense involves more diverse categories, ranging from accessaries to shoes, and thus are easier to be distinguished than FashionVC. 5) The performance of almost all methods can be strengthened with the hash code length increasing, which confirms the fact that more information regarding instances can be encoded by longer hash codes. Notably, if not declared, the hash code length is set as 128 for all the following experiments.

To gain more deep insights, we further investigated the performance of the proposed HiCHNet on difference categories. Here we chose DCMH as our baseline due to the fact that it also adopts the deep learning networks like ours and achieves overall satisfactory performance. As can be seen from Figure 4, our HiCHNet consistently shows superiority over DCMH across different categories on both datasets, proving the effectiveness of our model once again. Meanwhile, we find that the MAP scores of different categories are largely different. For example, in Figure 4(a), the performance of HiCHNet on "Short" is far better than that of "Activewear". One possible reason is that shorts are more visually distinctive than activewears.

5.4 On Component Analysis (RQ2)

To verify the effectiveness of each key component in our model, namely, the hierarchical discriminative learning and regularized cross-modal hashing, we investigated the nonnegative trade-off parameter γ in Eqn. (11). The sensitivity analysis of γ on Ssense is shown in Figure 5(a), where we varied γ from 0.1 to 1 with a step of 0.1. As can be seen, the optimal performance can be achieved when $\gamma = 0.5$, indicating that both components are essential to HiCHNet and their contributions are comparable.

As each layer of the hierarchy is assigned with the confidence in both components of hierarchical discriminative learning and regularized cross-modal hashing, we further studied the impact of the layer confidence on our HiCHNet. For simplicity, we unified the layer confidence for both components and set $\rho_1 = \tau_1$ and $\rho_2 = \tau_2$. We then changed ρ_1 from 0.1 to 0.9 with the step of 0.1, while



(a) Image→Text (b) Text→Image

Figure 6: Performance of HiCHNet and HiCHNet-flat on FashionVC.

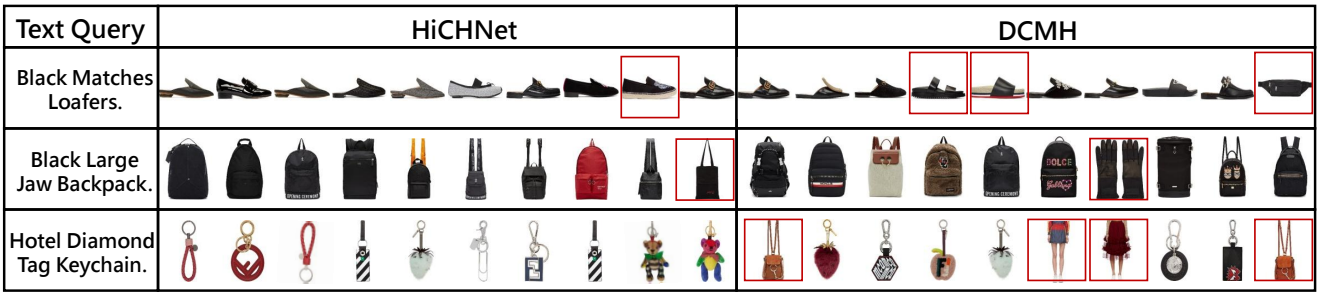


Figure 7: Illustration of ranking results from the whole retrieval set. The irrelevant images are highlighted in red boxes.

Figure 8: Illustration of ranking results from the constrained retrieval set.

keeping $\rho_1 + \rho_2 = 1$ and $\tau_1 + \tau_2 = 1$. Figure 5(b) shows the MAP curve with respect to coefficients ρ_1 and τ_1 on Ssense, where we fixed the value of parameter γ as 0.5 and code length as 128 bits. We observed that the best performance can be obtained by $\rho_1 = \tau_1 = 0.2$ (i.e., $\rho_2 = \tau_2 = 0.8$), which implies that both fine-grained and coarse-grained categories complementarily characterize the fashion items and contribute to the semantic similarity encoding. Moreover, this also reflects that fine-grained categories are more powerful than the coarse-grained ones, which is reasonable as fine-grained labels encode more detailed semantic information of the instance and provide more accurate instruction on the hash code learning.

5.5 On Label Hierarchy (RQ3)

To better explain the benefit of incorporating the label hierarchy especially in cross-modal hashing, we conducted the comparative experiment with one derivative of our model that does not take the hierarchy into consideration and only contain the K-th discriminative learning and the K-th regularized cross-modal hashing, termed as HiCHNet-flat. Figure 6 shows the performance of HiCHNet and HiCHNet-flat on FashionVC. As can be seen, HiCHNet consistently outperforms HiCHNet-flat no matter what the hash code length is set, and this well validates the necessity of taking into account the label hierarchy in the context of cross-modal hashing in fashion domain.

To thoroughly understand our model, we provided certain intuitive ranking results of HiCHNet and DCMH with two settings. On one hand, we listed the top 10 image results retrieved from the whole retrieval set in Figure 7, where the incorrect images are highlighted by red boxes. As can be seen, overall, our model is able to not only return fewer irrelevant images but also rank them at bottom positions as compared with DCMH, which confirms the superior performance of our model. On the other hand, towards more clear illustration, we replaced the whole retrieval set with a constrained subset of 10 images involving different categories and showed the retrieval results in Figure 8. As can be seen, HiCHNet outperforms DCMH again by ranking all the relevant results in the top places. Moreover, interestingly, we found that even for irrelevant instances, HiCHNet would rank them based on their semantic similarity to the given query irrelevance. For example, given the text query regarding category "Loafers", HiCHNet ranked "Boots" before "Dress" although both "Boots" and "Dress" are irrelevant to the query.

6 CONCLUSION AND FUTURE WORK

In this paper, we focus on studying the problem of cross-media retrieval with hierarchical categories, which has the great demand in fashion domain. We present a novel end-to-end supervised hierarchical cross-modal hashing method, consisting of two key components: the hierarchical discriminative learning and regularized cross-modal hashing. In addition, we constructed a benchmark dataset consisting of 15,696 image-text pairs from Ssense, labeled by 32 hierarchical categories. Extensive experiments have been conducted on real-world datasets and the results demonstrate the effectiveness of the proposed scheme and validate the benefits of utilizing the category hierarchy in cross-modal hashing. Interestingly, we found that the two key components comparably contribute to cross-modal hashing, which confirms the necessity of retaining the hierarchical discriminative capability of hash codes. Currently, we adopt the universal confidence for each layer. In future, we plan to adaptively assign the layer confidence for different instances to further improve the performance.

7 ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China, No.: 61702300, No.: 61772310, No.: 61702302, No.: 61802231, and No.: U1836216; the Project of Thousand Youth Talents 2016; the Tencent AI Lab Rhino-Bird Joint Research Program, No.: JR201805; the Future Talents Research Funds of Shandong University, No.: 2018WLJH63.

REFERENCES

[1] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. 2018. Cross-Modal Hamming Hashing. In *Europeon Conference on Computer Vision*. 207–223.

[2] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. 2017. Collective Deep Quantization for Efficient Cross-Modal Retrieval. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 3974–3980.

[3] Tiago Carvalho, Edmar R. S. De Rezende, Matheus T. P. Alves, Fernanda K. C. Balieiro, and Ricardo B. Sovat. 2017. Exposing Computer Generated Images by Eye's Region Classification via Transfer Learning of VGG19 CNN. In *16th IEEE International Conference on Machine Learning and Applications*. 866–870.

[4] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *British Machine Vision Conference*.

[5] Zhikui Chen, Fangming Zhong, Geyong Min, Yonglin Leng, and Yiming Ying. 2018. Supervised Intra- and Inter-Modality Similarity Preserving Hashing for Cross-Modal Retrieval. *IEEE Access* 6 (2018), 27796–27808.

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*.

[7] Guiguang Ding, Yuchen Guo, and Jile Zhou. 2014. Collective Matrix Factorization Hashing for Multimodal Data. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2083–2090.

[8] Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. 2016. Large-Scale Cross-Modality Search via Collective Matrix Factorization Hashing. *IEEE Transactions on Image Processing* 25, 11 (2016), 5427–5440.

[9] Fei Dong, Xiushan Nie, Xingbo Liu, Leilei Geng, and Qian Wang. 2018. Cross-modal Hashing Based on Category Structure Preserving. *Journal of Visual Communication and Image Representation* 57 (2018), 28–33.

[10] Yunchao Gong and Svetlana Lazebnik. 2011. Iterative quantization: A procrustean approach to learning binary codes. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*. 817–824.

[11] Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval*. 39–43.

[12] Go Irie, Hiroyuki Arai, and Yukinobu Taniguchi. 2015. Alternating Co-Quantization for Cross-Modal Hashing. In *IEEE International Conference on Computer Vision*. 1886–1894.

[13] Qing-Yuan Jiang and Wu-Jun Li. 2016. Deep Cross-Modal Hashing. *Computing Research Repository* abs/1602.02255 (2016).

[14] Liu Ke, Jun Wang, and Zhixian Ye. 2016. Fast-Gaussian SIFT for Fast and Accurate Feature Extraction. In *Advances in Multimedia Information Processing*. 355–365.

[15] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. 2018. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4242–4251.

[16] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable Multimodal Retrieval for Fashion Products. In *ACM Multimedia Conference on Multimedia Conference*. 1571–1579.

[17] Zijia Lin, Guiguang Ding, Jungong Han, and Jianmin Wang. 2017. Cross-View Retrieval via Probability-Based Semantics-Preserving Hashing. *IEEE Transactions on Cybernetics* 47, 12 (2017), 4342–4355.

[18] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. 2015. Semantics-preserving Hashing for Cross-view Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3864–3872.

[19] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2019. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning. *IEEE Transactions on Image Processing* 28, 3 (2019), 1235–1247.

[20] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *ACM Multimedia Conference on Multimedia Conference*. 843–851.

[21] Xin Liu, An Li, Ji-Xiang Du, Shu-Juan Peng, and Wentao Fan. 2018. Efficient Cross-modal Retrieval via Flexible Supervised Collective matrix factorization hashing. *Multimedia Tools and Applications* 77, 21 (2018), 28665–28683.

[22] Xu Lu, Lei Zhu, Zhiyong Cheng, Xuemeng Song, and Huaxiang Zhang. 2019. Efficient Discrete Latent Semantic Hashing for Scalable Cross-modal Retrieval. *Signal Processing* 154 (2019), 217–231.

[23] Lei Ma, Hongliang Li, Fanman Meng, Qingbo Wu, and King Ngi Ngan. 2018. Global and Local Semantics-Preserving Based Deep Hashing for Cross-modal retrieval. *Neurocomputing* 312 (2018), 49–62.

[24] Devraj Mandal, Kunal N. Chaudhury, and Soma Biswas. 2019. Generalized Semantic Preserving Hashing for Cross-Modal Retrieval. *IEEE Transactions on Image Processing* 28, 1 (2019), 102–112.

[25] Ralph Martinez, D. Smith, and H. Trevino. 1992. ImageNet: a global distributed database for color image storage, and retrieval in medical imaging systems. In *Fifth Annual IEEE Symposium on Computer-Based Medical Systems*. 710–719.

[26] Jonathan Masci, Michael M. Bronstein, Alexander M. Bronstein, and Jürgen Schmidhuber. 2014. Multimodal Similarity-Preserving Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 4 (2014), 824–830.

[27] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. *IEEE Transactions on Multimedia* 20, 2 (2018), 405–420.

[28] Dimitrios Rafailidis and Fabio Crestani. 2016. Cluster-based Joint Matrix Factorization Hashing for Cross-Modal Retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 781–784.

[29] Ali Ramezani-Kebrya, Ashish Khisti, and Ben Liang. 2018. On the Stability and Convergence of Stochastic Gradient Descent with Momentum. *Computing Research Repository* abs/1809.04564 (2018).

[30] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. 2013. Inter-media Hashing for Large-scale Retrieval from Heterogeneous Data sources. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 785–796.

[31] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. 2018. Neural Compatibility Modeling with Attentive Knowledge Distillation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 5–14.

[32] Xuemeng Song, Liqiang Nie, Luming Zhang, Maofu Liu, and Tat-Seng Chua. 2015. Interest Inference via Structure-Constrained Multi-Source Multi-Task Learning. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2371–2377.

[33] Mason Swofford. 2018. Image Completion on CIFAR-10. *Computing Research Repository* abs/1810.03213 (2018).

[34] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He. 2015. Semantic Topic Multimodal Hashing for Cross-Media Retrieval. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. 3890–3896.

[35] Dan Wang, Heyan Huang, Chi Lu, Bo-Si Feng, Guihua Wen, Liqiang Nie, and Xianling Mao. 2018. Supervised Deep Hashing for Hierarchical Labeled Data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 7388–7395.

[36] Lin Wu, Yang Wang, and Ling Shao. 2019. Cycle-Consistent Deep Generative Hashing for Cross-Modal Retrieval. *IEEE Transactions on Image Processing* 28, 4 (2019), 1602–1612.

[37] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning Discriminative Binary Codes for Large-scale Cross-modal Retrieval. *IEEE Transactions on Image Processing* 26, 5 (2017), 2494–2507.

[38] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, and Yueting Zhuang. 2014. Discriminative Coupled Dictionary Hashing for Fast Cross-media Retrieval. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 395–404.

[39] Dongqing Zhang and Wu-Jun Li. 2014. Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2177–2183.

[40] Lei Zhang, Yongdong Zhang, Richang Hong, and Qi Tian. 2015. Full-Space Local Topology Extraction for Cross-Modal Retrieval. *IEEE Transactions on Image Processing* 24, 7 (2015), 2212–2224.

[41] Peichao Zhang, Wei Zhang, Wu-Jun Li, and Minyi Guo. 2014. Supervised Hashing with Latent Factor Models. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 173–182.

[42] Yi Zhen and Dit-Yan Yeung. 2012. Co-Regularized Hashing for Multimodal Data. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems*. 1385–1393.

[43] Jile Zhou, Guiguang Ding, and Yuchen Guo. 2014. Latent Semantic Sparse Hashing for Cross-modal Similarity Search. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 415–424.

[44] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. 2013. Linear Cross-modal Hashing for Efficient Multimedia Search. In *ACM Multimedia Conference*. 143–152.