

# AN ADVERSARIAL AND DEEP HASHING-BASED HIERARCHICAL SUPERVISED CROSS-MODAL IMAGE AND TEXT RETRIEVAL ALGORITHM

Ruidong Chen,\* Baohua Qiang,\* Mingliang Zhou,\*\* Shihao Zhang,\* Hong Zheng,\* and Chenghua Tang\*

## Abstract

With the rapid development of robotics and sensor technology, vast amounts of valuable multimodal data are collected. It is extremely critical for a variety of robots performing automated tasks to find relevant multimodal information quickly and efficiently in large amounts of data. In this paper, we propose an adversarial and deep hashing-based hierarchical supervised cross-modal image and text retrieval algorithm to perform semantic analysis and association modelling on image and text by making full use of the rich semantic information of the label hierarchy. First, the modal adversarial block and the modal differentiation network both perform adversarial learning to keep different modalities with the same semantics closest to each other in a common subspace. Second, the intra-label layer similarity loss and inter-label layer correlation loss are used to fully exploit the intrinsic similarity existing in each label layer and the correlation existing between label layers. Finally, an objective function for different semantic data is redesigned to keep data with different semantics away from each other in a common subspace, thus avoiding interference of retrieval by data of different semantics. The experimental results on two cross-modal retrieval datasets with hierarchically supervised information show that the proposed method substantially enhances retrieval performance and consistently outperforms other state-of-the-art methods.

## Key Words

Cross-modal image and text retrieval, deep hash algorithm, hierarchical supervision, adversarial network

## 1. Introduction

In recent years, various types of intelligent robots [1], [2] have developed rapidly, *e.g.*, clothing guide

robots. Cross-modal retrieval [3], [4], a key technology for robots to achieve automated tasks [5], [6] through the understanding of multimodal content, is the process of retrieving data from one modality and returning data from other modalities that are most semantically relevant to the retrieved data. For example, if users are visiting a clothing shop, by submitting a photo of their favourite, they can get the relevant image and text details simultaneously.

In recent years, many approaches are proposed to address cross-modal retrieval. Hardoon *et al.* [7] proposed the canonical correlation analysis (CCA) method to map different modalities into a shared subspace and maximise the correlation of different modalities sharing the same semantic information by using statistical analysis. In the storage and retrieval of large-scale cross-modal data, hashing algorithms are widely regarded for their low-storage cost and high-retrieval efficiency. Jang *et al.* [8] proposed a deep cross-modal hashing (DCMH) method to integrate feature learning and hash code learning into a unified framework. Li *et al.* [9] proposed a self-supervised adversarial hashing (SSAH) method to build self-supervised semantic networks by using labels as self-supervised information.

Most of the existing cross-modal retrieval methods are used for nonhierarchically structured supervised data, and cannot fully exploit the supervised information of the labels. However, in many real-world application scenarios, label-supervised information on cross-modal data often has some kind of hierarchical structure with rich semantic information. For example, in the field of public security, the image or video automatically collected by robots through sensors may contain multiple layers of label supervision information.

Currently, there are only a few methods that have been designed to label supervision information in hierarchical structures. Wang *et al.* [10] proposed the supervised hierarchical deep hashing (SHDH) method, which defines a similarity formula to weight different levels for labelled supervised information of the hierarchy and verifies that the hierarchical information can improve the hash retrieval

\* Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China; e-mail: pgezcrb@163.com; qiangbh@guet.edu.cn; shihaozhang2022@163.com; 458436419@qq.com; tch@guet.edu.cn

\*\* School of Computer Science, Chongqing University, Chongqing 400044, China; e-mail: mingliangzhou@cqu.edu.cn  
Corresponding author: Baohua Qiang

accuracy. However, this method is designed for single-modal retrieval. To verify the effectiveness of labels with hierarchical structure in cross-modal retrieval, Sun *et al.* [11] proposed the supervised hierarchical cross-modal hashing (HiCHNet) method to learn hierarchical information and regularised cross-modal hashing simultaneously. However, those methods have the following problems.

- The distance between multimodal data with the same semantic information in the common subspace is not sufficiently minimised.
- The inter-layer correlation of supervisory information is not sufficiently considered, so that complex inter-layer correlation information is not fully learnt.
- Cross-modal retrieval has been interfered by dissimilar data.

To address the above problems, we propose a novel method for hierarchical supervised cross-modal image and text retrieval. The contributions of this study are as follows.

- The feature extraction network and the modality differentiation network, which are used as generators and adversaries, respectively, both perform adversarial learning to result in the closest distance in the common space for different modalities containing the same semantics.
- The intra-label layer similarity loss and inter-label layer correlation loss are introduced to fully explore the intrinsic similarity existing in each layer of labels and the correlation existing between label layers, thus improving the accuracy of cross-modal retrieval.
- An objective function for the distance between different semantic categories of data is redesigned to keep the modal data of different semantic categories distant from each other in the common space.

The remainder of this paper is organised as follows. Related work is reviewed in Section 2. The proposed algorithm is illustrated in Section 3. The experimental results for two cross-modal datasets with hierarchical supervised information are presented in Section 4. Finally, conclusions are drawn in Section 5.

## 2. Related Work

The traditional cross-modal retrieval methods [7], [12]–[14] construct a matrix for different media, projecting it uniformly into a shared subspace, and then utilise distance metrics, such as Euclidean distance or cosine distance, to measure the similarity between heterogeneous modalities. CCA [7] is widely used in cross-modal retrieval, and many cross-modal retrieval methods have been built on it. However, the problem of the “heterogeneity gap” is still not effectively addressed by most traditional cross-modal retrieval methods which rely on hand-designed features.

Deep neural networks have made progress in many fields, such as computer vision [15], [16] and natural language processing [17], [18] and have also been effectively adopted in cross-modal retrieval. However, there are problems of high storage costs and slow retrieval speed

when employing the deep learning methods [19]–[21] for cross-modal retrieval of large-scale data.

In the storage and retrieval of large-scale cross-modal data, hashing algorithms [22]–[26] are widely regarded for their low storage cost and high retrieval efficiency. Jiang *et al.* [8] proposed DCMH to integrate feature learning and hash code learning into a unified framework. Li *et al.* [9] proposed the SSAH method to build self-supervised semantic networks by using labels as self-supervised information.

At present, there is only a little work on cross-modal data for multilayer label supervision. Sun *et al.* [11] proposed the HiCHNet method to effectively utilise label hierarchy information for facilitating hash code learning through hierarchical discriminative learning and regularised cross-modal hashing methods. However, the distance between multimodal with the same semantic information in the common subspace is not sufficiently minimised, the inter-layer correlation of supervisory information is not sufficiently considered, and the different semantic data are not sufficiently separated.

## 3. An Adversarial and Deep Hashing-based Hierarchical Supervised Cross-modal Image and Text Retrieval Algorithm

Dataset  $\beta(i) = \{(x^i, t^i) \mid i \in 1, 2, \dots, N\}$  has  $N$  sets of image-text pairs, where  $x^i$  is the original feature vector of the  $i$ -th image data,  $p_x$  is the feature dimension of  $x^i$ ,  $t^i$  is the representation of the  $i$ -th text data, and  $q_t$  is the feature dimension of  $t^i$ . The image-text pair is labelled by  $E$  layers, with the label layer indexed from top to bottom as  $\{1, 2, \dots, E\}$ , and  $\Phi_e$  denotes the total number of labels in the  $e$ -th layer. There is a label vector  $S_e = \{s_i^e\}_{e=1}^E$  for each set of image-text pairs  $\beta(i)$ , where  $s_i^e = \{s_i^{e1}, s_i^{e2}, \dots, s_i^{e\Phi_e}\}$ , and  $s_i^{ej} = 1$  denotes that the  $i$ -th image-text pair data is labelled by the  $j$ -th label of the  $e$ -th layer; otherwise,  $s_i^{ej} = 0$ .

### 3.1 Overview

As shown in Fig. 1, our method consists of three blocks. The feature extraction block consists of an image feature extraction task and a text feature extraction task. Each image in the dataset is resized to  $224 \times 224$  and fed into a deep neural network to extract the high-dimensional features of the image. The deep neural network is a VGG-16 pre-trained on the large-scale dataset ImageNet [27]. The last layer of the original network structure was modified to be the output layer of the hash code, the number of neurons is the hash code length value, and the output is mapped to between  $-1$  and  $1$  using the Tanh activation function. Text feature extraction consists of three steps. First, a bag-of-words (BOW) model is used to represent each text, and a BOW model of the text modality is used as the input to the network. Then, the high and low-level features under different perceptual fields are extracted by a multiscale feature stacking model (MSFSM) constructed from five parallel levels of mean pooling layers with window sizes of  $1 \times 1$ ,  $2 \times 2$ ,  $3 \times 3$ ,  $5 \times 5$ , and  $10 \times 10$ , respectively.

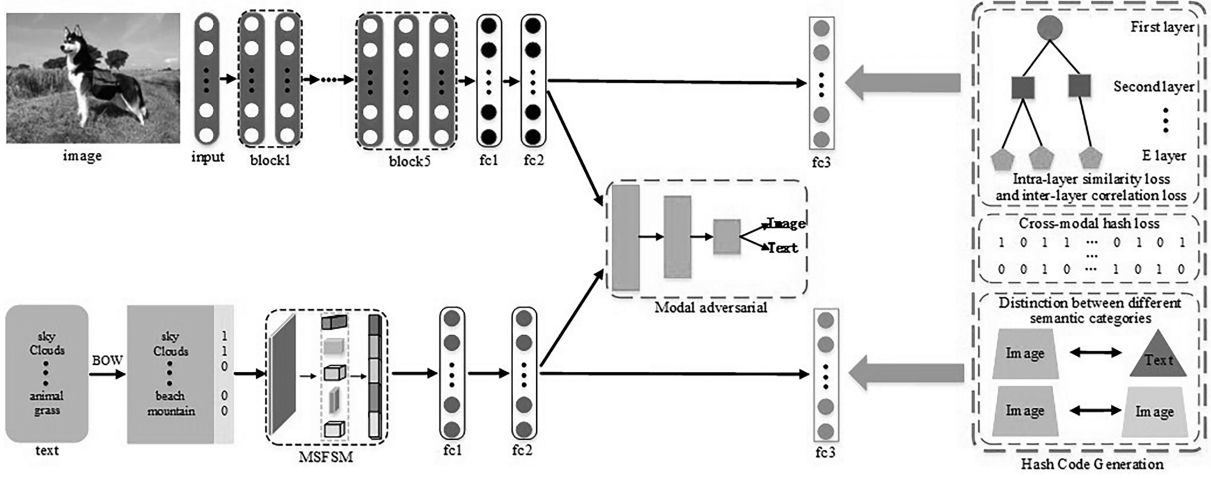


Figure 1. Flowchart of the proposed method.

Finally, the features are extracted by a neural network consisting of three fully connected layers and then a hash code of the text features is output.

In the modal adversarial block, the feature extraction network and the modal differentiation networks both perform adversarial learning until it is difficult for modal differentiation networks to distinguish the modal type of the modal features extracted by the feature extraction network, which results in the closest distance in the common space between different modal data that share the same semantics.

The hash code generation block contains four branches, namely, cross-modal hash loss, intra-label layer similarity loss, inter-label layer relevance loss, and different semantic class differentiation. Cross-modal hash loss is used to enable the model to perform both feature learning and hash code learning simultaneously. It is extremely essential to fully exploit the interconnections within the label hierarchy. Therefore, intra-label layer similarity loss and inter-label layer correlation loss are used to fully exploit the intrinsic similarity of each layer of labels and the inter-label layer correlation. The different objective functions for the same modal and different modal data of different semantic categories are set separately, so that the modal of different semantic categories are kept at a certain distance from each other in the common space, thus avoiding the interference of cross-modal retrieval by data of different semantic.

### 3.2 Modal Adversarial Block

The output of the fc2 layer of the image feature extraction block and the text feature extraction block is fed into the modal adversarial block. Modal adversarial is based on the adversarial idea of learning the common space of image modality and text modality. The task of the feature extraction is representation learning of image and text modalities, mapping image and text modalities into a common subspace, to confuse the discrimination as an adversary of the modal adversarial block, and thus improving the discriminatory power of the adversarial

block. The task of the modal adversarial block is to discriminate the modal type of the samples in the feature extraction block to enhance the representational power of the feature extraction block and further minimise the distance between different modal data with the same semantic information in a common subspace.

The  $E$  networks with three fully connected layers are used by the modal adversarial block, with the first hidden layer having the same number of nodes as the input feature dimension, the second hidden layer having the same number of nodes as the total number of labels  $\Phi_e$  in layer  $e$ , and the third layer having a node count of 2. Its activation function is a sigmoid function, and the output is a binary code with 0 indicating image modality and 1 indicating text modality.

The cross-entropy loss function used in the modal adversarial block is defined as follows:

$$L_{Adv} = \sum_{e=1}^E L_{Adv,e} \quad (1)$$

$$L_{Adv,e}(\varepsilon) = -\frac{1}{N} \sum_{i=1}^N \left( v_i \cdot (\log G(x^i; \varepsilon) + \log(1 - G(t^i; \varepsilon))) \right) \quad (2)$$

where  $L_{Adv}$  denotes the overall objective function of the modal adversarial block,  $L_{Adv,e}$  denotes the adversarial loss corresponding to the layer  $e$  labels.  $v_i$  denotes the true label supervision information for each data, and  $G(*; \varepsilon)$  denotes the modal probability distribution generated by the sample  $\beta(i)$  in the modal adversarial network.  $\varepsilon$  is the parameter of the modal adversarial block.

### 3.3 Hash Code Generation Block

#### 3.3.1 Cross-modal Hashing

The output of the feature extraction block is introduced into the fc3 layer and the objective function of the cross-modal hash is defined as follows:

$$L_{hash} = \|D_x - f(x^i, W_f)\|_F^2 + \|D_t - g(t^i, W_g)\|_F^2 \quad (3)$$

where  $f(x^i, W_f)$  and  $g(t^i, W_g)$  denote the image and text features of the sample  $\beta_i$  learnt by the feature extraction block, respectively.  $W_f$  and  $W_g$  denote the network parameters of the image and text modal, respectively.  $D_x \in \{-1, 1\}^h$  and  $D_t \in \{-1, 1\}^h$  are the hash codes learnt from the image and text modalities, respectively.  $\|\bullet\|_F$  represents the Frobenius norm. The neural network parameters and the binary hash codes are learnt in the same objective function.

### 3.3.2 Intra-label Similarity

To maintain the similarity of labels at each layer in the label hierarchy, a label hash code  $C_e \in \{-1, 1\}^{h \times \Phi_e}$ ,  $e \in \{1, 2, \dots, E\}$  is generated for each label at each layer. The loss function is as follows:

$$L_{\text{intra\_layer}} = \sum_{e=1}^E \zeta_e \left( \|hS_e - f(x^i, W_f) C_e\|_F^2 + \|hS_e - g(t^i, W_g) C_e\|_F^2 \right) \quad (4)$$

where  $C_e$  is the class hash code at the  $e$ -th layer,  $\zeta_e$  is the confidence degree of the labels at the  $e$ -th layer, and the sum of the confidence degrees of all label layers is 1.

### 3.3.3 Inter-label Correlation

It is extremely important to mine the relevance of different layers in a hierarchy of labels. To fully capture cross-layer relevance, a cross-layer label similarity matrix is defined as follows:

$$\Psi_{eE}^{ij} = \begin{cases} 1, & \text{if } E^j \text{ is a descendant node of } e^i \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where  $e \in \{1, 2, \dots, E-1\}$ .  $e^i$  represents the  $i$ -th label of the  $e$ -th layer and  $E$  represents the last layer label.  $\Psi_{eE}^{ij} = 1$  denotes that the  $j$ -th label of the  $E$ -layer is a descendant node of the  $i$ -th label of the  $e$ -th layer; otherwise,  $\Psi_{eE}^{ij} = 0$ . The objective function for the cross-layer association of labels is as follows:

$$L_{\text{inter\_layer}} = \sum_{e=1}^E \eta_e \left( \|h\Psi_{eE} - C_E C_e\|_F^2 \right) \quad (6)$$

where  $C_E$  is the hash matrix of the  $E$ -layer labels,  $\eta_e$  represents the hyperparameters, and  $\sum_{e=1}^{E-1} \eta_e = 1$ .

### 3.4 Different Semantic Classes Distinguish Blocks

In the common subspace, the distance between different semantic categories and different modalities should be as large as possible. The deep neural network extracts feature for the  $i$ -th image text pair and the  $j$ -th image text pair and maps them into a common subspace using a hash function. The image features are  $F(x^i)$  and  $F(x^j)$ , and the text features are  $G(t^i)$  and  $G(t^j)$ .

The objective function for the distance between different semantic classes and different modal data can be

written as:

$$L_1(i, j) = \left| \text{dis}(F(x^i), G(t^j)) - \sqrt{\text{dis}(F(x^i), F(x^j)) \bullet \text{dis}(G(t^i), G(t^j))} \right| + \left| \text{dis}(F(x^j), G(t^i)) - \sqrt{\text{dis}(F(x^i), F(x^j)) \bullet \text{dis}(G(t^i), G(t^j))} \right| \quad (7)$$

where  $\text{dis}(W, V)$  denotes the cosine distance between  $W = (w_1, w_2, \dots, w_n)$  and  $V = (v_1, v_2, \dots, v_n)$ .

Identical modal data of different semantic categories should also be separated in the common subspace. The objective function is as follows:

$$L_2(i, j) = \left| \text{dis}(F(x^i), F(x^j)) - \sqrt{\text{dis}(F(x^i), F(x^j)) \bullet \text{dis}(G(t^i), G(t^j))} \right| + \left| \text{dis}(G(t^i), G(t^j)) - \sqrt{\text{dis}(F(x^i), F(x^j)) \bullet \text{dis}(G(t^i), G(t^j))} \right| \quad (8)$$

The total objective function for the different semantic classes of distinguished blocks is defined as follows:

$$L_{\text{dis}} = \sum_{i,j=1}^N (L_1(i, j) + L_2(i, j)) \quad (9)$$

The overall loss function is defined as follows:

$$\text{Loss} = \alpha L_{\text{Adv}} + \beta L_{\text{hash}} + \chi L_{\text{intra\_layer}} + \delta L_{\text{inter\_layer}} + \phi L_{\text{dis}} \quad (10)$$

where  $\alpha, \beta, \chi, \delta$  and  $\phi$  are hyperparameters.

## 3.5 Summary of Our Method

Our method is summarised in Algorithm 1. Our method has four parameters that need to be trained and optimised, which are network parameter  $W_f$  for extracting image features, network parameter  $W_g$  for extracting text features, and hash codes  $D_x$  and  $D_t$ . First, we randomly select the mini-batch of image and text data from  $\beta(i)$ . Second,  $f(x^i, W_f)$  and  $g(t^i, W_g)$  are calculated using a forward propagation network, and parameters  $W_f$  and  $W_g$  are updated using the Adam gradient descent algorithm in backward propagation. Finally, the hash codes  $D_x$  and  $D_t$  are updated using (3) in the hash code generation block. The hyperparameters  $\alpha, \beta, \chi, \delta$ , and  $\phi$  are given values by the grid search method to make the model optimal. Our method will perform both feature learning and hash code learning, enabling end-to-end learning.

## 4. Experiment

### 4.1 Experimental Setting

We used two datasets with hierarchically supervised information for cross-modal image and text retrieval: the FashionVC dataset [28] and the Ssense dataset [11]. The

**Algorithm 1** Optimisation process of our algorithm.

<b>Input</b>	Image-text Training set: $\beta(i) = \{(x^i, t^i) \mid i = 1, 2, \dots, N\}$ , where $x^i$ denotes the image data and $t^i$ denotes the text data. Label vectors: $S_e = \{s_e^e\}_{e=1}^E$ . The length of the hash code: $h$ . Batch size: mini-batch. Maximum iterations: $T_{\max} = \lceil N/\text{mini-batch} \rceil$ . Hyperparameters: $\alpha, \beta, \chi, \delta$ , and $\phi$ .
<b>Output</b>	Hash Codes $D_x$ and $D_t$ ; parameters $W_f$ of the image feature extraction network; parameters $W_g$ of the text feature extraction network.
1:	<b>Initialisation:</b> Initialise all parameters in the model.
2:	<b>Repeat</b>
3:	<b>for</b> item = 1 to $T_{\max}$ <b>do</b>
4:	Randomly select the mini-batch of image data from $\beta(i)$ .
5:	Calculate $f(x^i, W_f)$ using forward propagation.
6:	Update the parameters using Adam gradient descent algorithm: $W_f = \text{Adam}(\nabla W_f(f(x^i, W_f)), W_f)$ .
7:	Randomly select the mini-batch of text data from $\beta(i)$ .
8:	Calculate $g(t^i, W_g)$ using forward propagation.
9:	Update the parameters using Adam gradient descent algorithm: $W_g = \text{Adam}(\nabla W_g(g(t^i, W_g)), W_g)$ .
10:	<b>end for</b>
11:	Calculate equation (3) updating hash codes $D_x = \text{sign}(f(x^i, W_f))$ and $D_t = \text{sign}(g(t^i, W_g))$ , where $\text{sign}(\bullet)$ is a symbolic function [8].
12:	<b>until</b> fixed number of iterations $T_{\max}$ or achieved convergence.

Mean Average Precision (MAP) [29] and the Precision-Recall curve (PR curve) were used to measure the performance of the model.

#### 4.2 Comparison of State-of-the-art Methods

Our method was compared on two retrieval tasks: retrieving text by image (I2T) and retrieving image by text (T2I) with four state-of-the-art methods, including an unsupervised method: CCA [7], and three supervised methods: DCMH [8], SSAH [9], and HiCHNet [11], where HiCHNet is a labelling hierarchy-based approach. There are two different variants of the method for the three non-hierarchical label structures, depending on the way the labels are entered. The first variant combines the labels of all layers in the dataset to form a complete label and is entered into the nonhierarchical method with a suffix marked with “-a”. The second variant is tagged with the second level tag only and then entered into the nonhierarchical method with the suffix “-s” tag.

Table 1 shows the MAP values of the proposed method and the comparison method for different hash code lengths on the FashionVC and Ssense datasets. Bolded fonts are the best MAP values for the comparison methods, and underlined are the second-best MAP values. The conclusions are as follows.

The experimental results show that our method significantly outperforms other compared methods on both the FashionVC dataset and Ssense dataset for two retrieval tasks with different hash code lengths. For

example, comparing the state-of-the-art method with the proposed method in the I2T tasks on the FashionVC dataset, the MAP improved by 14.5%, 11.5%, 8.2%, and 8.8% for hash code lengths of 16, 32, 64, and 128 bits, respectively.

Both our method and HiCHNet are designed for hierarchical labelling and perform better than other compared methods. The importance of considering label hierarchies in supervised cross-modal hash retrieval is fully confirmed.

The MAP values for methods using all labels are mostly lower than for methods using only a second layer of labels. The possible reasons are that the method using all labels simply combines all labels and ignores the hierarchical relationship of the labels. The second level of labels is a more refined supervised representation of the modal data and so corresponds to a higher MAP value.

With an increasing number of hash code bits, other methods also obtained excellent MAP values. The possible reason is that the smaller number of hash code bits loses some of the semantic information and thus makes lower retrieval results.

In addition, we show the PR curves of various methods for two retrieval tasks with 16 and 128 bits on the two datasets, as shown in Fig. 2. The experimental results show that our method outperforms the four comparison methods in terms of precision and recall for two retrieval tasks on two datasets. Note that CCA, DCMH, and SSAH only use the second layer of label markings.

Table 1  
Performance Comparison Between Our Method and Other State-of-the-art Methods in Terms of MAP Values on the FashionVC and Ssense Datasets

Method	FashionVC								Ssense							
	I2T				T2I				I2T				T2I			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
CCA-s [7]	0.248	0.237	0.224	0.175	0.257	0.251	0.249	0.229	0.462	0.498	0.402	0.393	0.527	0.574	0.475	0.396
CCA-a [7]	0.226	0.213	0.208	0.169	0.238	0.229	0.217	0.211	0.457	0.493	0.391	0.311	0.512	0.561	0.462	0.351
DCMH-s [8]	0.509	0.629	0.637	0.667	0.632	0.685	0.701	0.729	0.648	0.687	0.698	0.736	0.642	0.657	0.712	0.748
DCMH-a [8]	0.503	0.587	0.612	0.631	0.607	0.657	0.691	0.708	0.613	0.651	0.676	0.721	0.609	0.632	0.684	0.729
SSAH-s [9]	<u>0.621</u>	<u>0.698</u>	0.702	0.426	0.723	0.782	0.797	0.433	0.447	0.456	0.307	0.274	0.442	0.367	0.236	0.127
SSAH-a [9]	0.610	0.663	0.693	0.398	0.729	0.805	0.816	0.472	0.546	0.604	0.639	0.387	0.457	0.465	0.329	0.278
HiCHNet [11]	0.613	0.689	<u>0.720</u>	<u>0.719</u>	<u>0.820</u>	<u>0.874</u>	<u>0.884</u>	<u>0.886</u>	<u>0.703</u>	<u>0.822</u>	<u>0.880</u>	<u>0.892</u>	<u>0.685</u>	<u>0.838</u>	<u>0.874</u>	<u>0.916</u>
<b>Our</b>	<b>0.711</b>	<b>0.778</b>	<b>0.779</b>	<b>0.782</b>	<b>0.905</b>	<b>0.936</b>	<b>0.941</b>	<b>0.947</b>	<b>0.915</b>	<b>0.942</b>	<b>0.945</b>	<b>0.943</b>	<b>0.937</b>	<b>0.956</b>	<b>0.959</b>	<b>0.961</b>

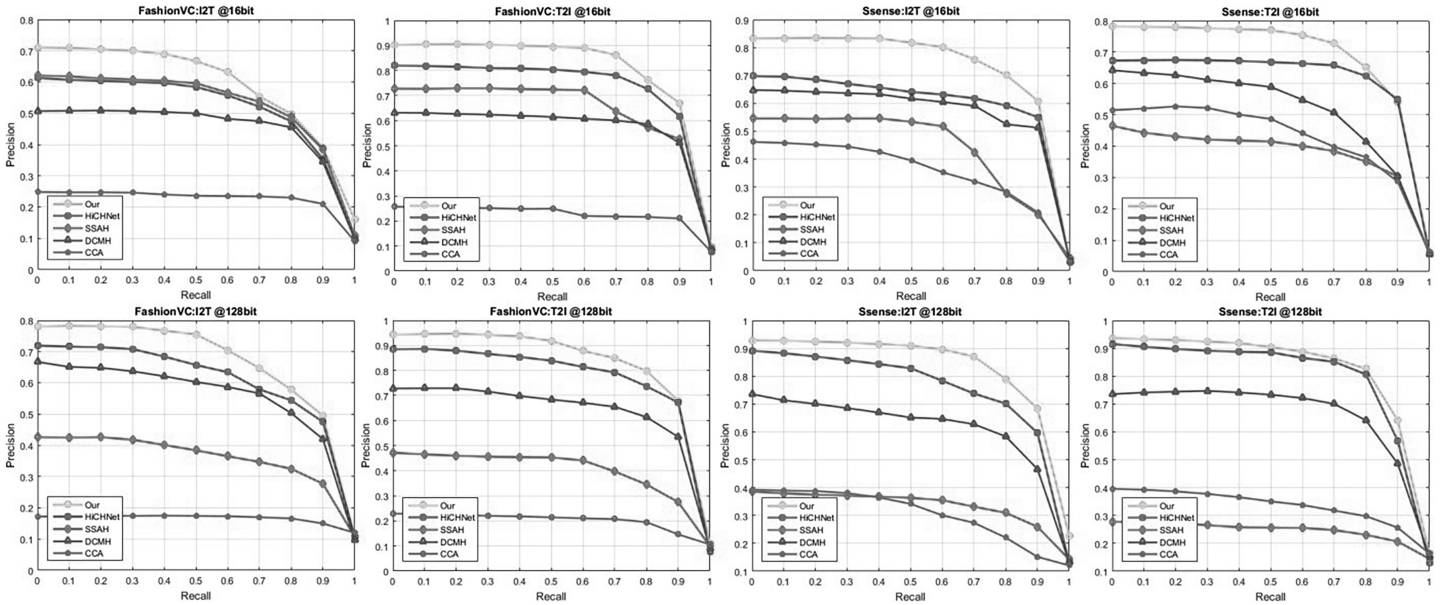


Figure 2. PR curves for the different methods with 16 and 128 bits on the two datasets.

We also report the actual display results for the proposed method and HiCHNet in Fig. 3, from which we can see that the proposed algorithm obtained a high retrieval accuracy than HiCHNet. Incorrect retrieval results of the HiCHNet caused by the fact that images of different categories are less distinguishable. For example, in the I2T task, the query label is "Outerwear>Jacket", but the error retrieval result of the HiCHNet is labelled "Outerwear>Coat". Extensive actual display results show that the proposed algorithm is most effective.

### 4.3 Further Analysis

To verify the effectiveness of our approach, we designed two variants. Var-1 based on our method to remove MSFSM and different semantic categories distinguish blocks. Var-2 is based on our approach of only removing different semantic categories distinguish blocks. Table 2 shows the MAP values for our method and the variant method on

the FashionVC and Ssense datasets. Bolded fonts are the best MAP values and underlined are the second best MAP values. The conclusions are as follows.

The main reason why Var-2 significantly outperforms Var-1 is that MSFSM effectively extracts features from low to high level and stacks them, greatly enriching the proposed features, and is a text feature extraction method that considers the semantics of the text.

Our method significantly outperforms other variants on both the FashionVC dataset and Ssense dataset for two retrieval tasks with different hash code lengths. For example, comparing the Var-2 with our method in the I2T tasks on the FashionVC dataset, the MAP improved by 1.0%, 0.6%, 0.6%, and 0.4% for hash code lengths of 16, 32, 64, and 128 bits, respectively. This demonstrates the need to consider the separation of different semantic categories of data and also confirms the contribution of MSFSM to cross-modal retrieval results.

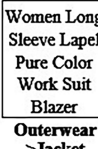

Query	Method	Results									
<div>  <p>Women Long Sleeve Lapel Pure Color Work Suit Blazer Outerwear &gt; Jacket</p> </div>	Our										
	HiCHNet										
<div>  <p>Dresses &gt; Day Dress</p> </div>	Our	Embroidered Lace Knee Length Dress Pink	Floral Sleeveless Ruffled Dress	Floral Embroidered Sweatshirt Dress	Blue Plaid Shirt Dress	Rubie Black Satin Embroidered Cami Dress	A Line Wrap Slip Dress Purple blue	T-Shirt Mock Dress Blue Polka	Embroidered lace pencil skirt	Khaki Womens Striped Sweater Dress	V-neck Short Sleeve Floral Print Midi Dress
	HiCHNet	Yoins High Neck Long Sleeves Knit Casual Dress	Blue Plaid Shirt Dress	Embroidered Lace Knee Length Dress Pink	A Line Wrap Slip Dress Purple blue	T-Shirt Mock Dress Pink Check	Peter Pilotto Embroidered Lace Top	Floral Sleeveless Ruffled Dress	Heart Patterned High Waist Knee Length Dress	Pink Tartan Plaid Coat	Yoins Plum Long Sleeves Sweater Dress

Figure 3. Example of cross-modal image and text retrieval by the proposed method and HiCHNet. Green boxes indicate correct retrieval results, and red boxes represent incorrect retrieval results.

Table 2  
MAP Comparison Between Our Method and Variants on the FashionVC and Ssense Datasets

Method	FashionVC								Ssense							
	I2T				T2I				I2T				T2I			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
Var-1	0.679	0.754	0.759	0.761	0.850	0.912	0.923	0.929	0.734	0.837	0.896	0.897	0.742	0.857	0.882	0.918
Var-2	<u>0.704</u>	<u>0.773</u>	<u>0.774</u>	<u>0.779</u>	<u>0.897</u>	<u>0.932</u>	<u>0.936</u>	<u>0.941</u>	<u>0.825</u>	<u>0.898</u>	<u>0.921</u>	<u>0.927</u>	<u>0.826</u>	<u>0.909</u>	<u>0.920</u>	<u>0.934</u>
Our	<b>0.711</b>	<b>0.778</b>	<b>0.779</b>	<b>0.782</b>	<b>0.905</b>	<b>0.936</b>	<b>0.941</b>	<b>0.947</b>	<b>0.915</b>	<b>0.942</b>	<b>0.945</b>	<b>0.943</b>	<b>0.937</b>	<b>0.956</b>	<b>0.959</b>	<b>0.961</b>

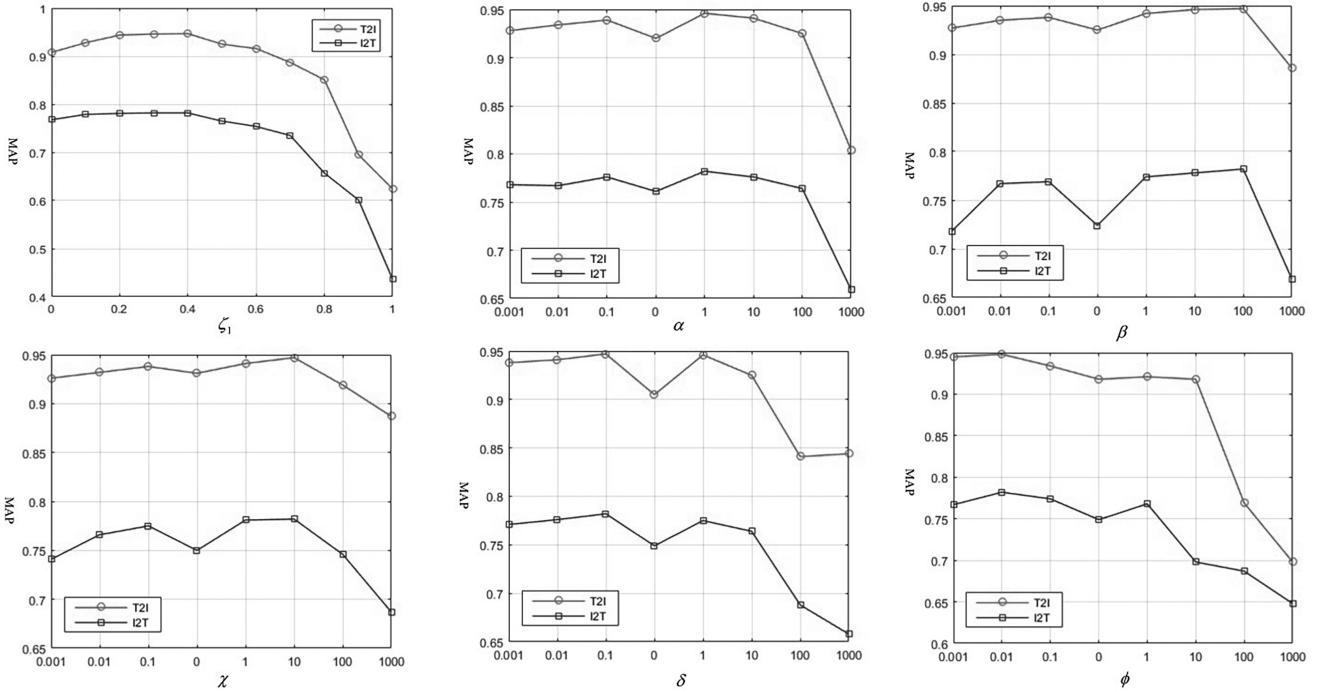


Figure 4. Plots of the MAP values for confidence degrees and hyperparameters of the proposed algorithm.

#### 4.4 Sensitivity of the Parameters

We explored the effect of confidence degrees and five hyperparameters on the MAP of two retrieval tasks in

cross-modal retrieval on the FashionVC dataset with a hash code of 128 bits. The FashionVC dataset has only two levels of class supervision information and  $\sum_{e=1}^{E-1} \eta_e = 1$ , so  $\eta_1 = 1$ . Fig. 4 shows a plot of the MAP values for

confidence degrees and hyperparameters. The FashionVC dataset has two layers of label supervision information ( $K = 2$ ) and  $\zeta_1 + \zeta_2 = 1$ . The highest MAP values for both the I2T and T2I tasks were achieved when  $\zeta_1$  in the range  $[0.2, 0.4]$ . The reason why  $\zeta_2$  is greater than  $\zeta_1$ 's the effect on MAP values, meaning that the second layer of labels is more important to retrieval accuracy than the first layer of labels, is that the second layer of labels is a more refined supervised representation of the modal data.

In addition, the optimal values for the five hyperparameters  $\alpha$ ,  $\beta$ ,  $\chi$ ,  $\delta$ , and  $\phi$  are searched by a grid search method. When any of the five hyperparameters equals 0, their MAP values are low, which fully illustrates the necessity of considering modal adversarial block loss, cross-modal hash loss, intra-label layer similarity loss, label cross-layer correlation loss, and different semantic category differentiation loss.

## 5. Conclusion

In this paper, we proposed an adversarial and deep hashing-based hierarchical supervised cross-modal image and text retrieval method that enables the robot to automatically understand and correlate key elements between different modal data, and to achieve relatively accurate cross-matching. Modal adversarial networks are introduced to reduce the distance between different modalities with the same semantics in a shared subspace. Intra-label similarity loss and inter-label correlation loss are also introduced to fully exploit the intrinsic similarity of each label layer and the inter-label correlation. Furthermore, an objective function for different semantic data is redesigned to keep the modal of different semantics at a certain distance from each other in the common space. Experimental results on two cross-modal retrieval datasets with hierarchical supervised information show that our method improves the MAP by 10.75%/14.48% and 7.70%/16.38% on average compared to the optimal comparison method on two tasks (I2T/I2T), which fully demonstrates its effectiveness in large-scale cross-modal retrieval with hierarchical supervised information. Meanwhile, the proposed algorithm outperforms the state-of-the-art methods significantly on the actual display results further verifying the effectiveness of the proposed algorithm.

In future work, a unified deep model that can learn multiple modals (*e.g.*, video, audio, and 3D models) simultaneously needs to be further investigated for intelligent robots. Furthermore, the proposed algorithm will be deployed on the clothing guide robot. The customer simply enters visual or textual information into the robot window and the robot retrieves the clothing information using the proposed algorithm and presents it to the customer. The proposed algorithm was trained on the fashion domain and we are collecting clothing image-text cross-modal data from clothing shops to further improve the generalisation of the algorithm, which will then be refined for deployment on clothing guide robots.

## Acknowledgement

This work was supported in part by the Natural Science Foundation of Guangxi under Grants 2019GXNSFDA185006 and 2019GXNSFDA185007; in part by the National Natural Science Foundation of China under Grants 62262006 and 62062028; in part by the Guangxi Key Research and Development Program under Grants AB23026048, AB17195053, and AD18281002; in part by the Guilin Science and Technology Development Program under Grants 20210104-1 and 20220115-1; and in part by the Innovation Project of GUET Graduate Education under Grant 2023YCX040.

## References

- [1] C. Shihuan, L. Wanlin, W. Shangsheng, H. Mouxiao, C. Ruihong, and G. Weipeng, Indoor localization system of ROS mobile robot based on visible light communication, *International Journal of Robotics & Automation*, 38(1), 2023, 1–12.
- [2] Z. Weiyang, Y. Jianjun, C. Houru, G. Lianxin, and Z. Wei, Human back acupuncture points location using RGB-D image for TCM massage robots, *International Journal of Robotics & Automation*, 38(1), 2023, 67–75.
- [3] S. Chun, S. J. Oh, R. S. de Rezende, *et al.*, Probabilistic embeddings for cross-modal retrieval, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Virtual Event*, 2021, 8415–8424.
- [4] W. Xie, M. Cui, M. Liu, P. Wang, and B. Qiang, Deep hashing multi-label image retrieval with attention mechanism, *International Journal of Robotics & Automation*, 37(4), 2022, 372–381.
- [5] Y. Feng, T. Tang, S. Chen, and Y. Wu, Automated defect detection based on transfer learning and deep convolution generative adversarial networks, *International Journal of Robotics & Automation*, 36(6), 2021, 471–478.
- [6] H. Liu, S. Ren, D. Ren, and X. Liu, Automatic extraction of orchards from remote sensing image based on category attention mechanism, *International Journal of Robotics & Automation*, 37(1), 2022, 20–28.
- [7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, *Neural Computation*, 16(12), 2004, 2639–2664.
- [8] Q. Y. Jiang, and W. J. Li, Deep cross-modal hashing, *Proc. 30th IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, 3270–3278.
- [9] C. Li, C. Deng, and N. Li, Self-supervised adversarial hashing networks for cross-modal retrieval, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, 4242–4251.
- [10] D. Wang, H. Huang, C. Lu, B.S. Feng, G. Wen, L. Nie, and X.L. Mao, Supervised deep hashing for hierarchical labeled data, *Proc. 32th AAAI Conf. on Artificial Intelligence*, New Orleans, LA, 2018, 7388–7395.
- [11] C. Sun, X. Song, F. Feng, W.X. Zhao, H. Zhang, and L. Nie, Supervised hierarchical cross-modal hashing, *Proc. 42nd International ACM Sigir Conf. on Research and Development in Information Retrieval*, Paris, France, 2019, 725–734.
- [12] X. Zhai, Y. Peng, and J. Xiao, Learning cross-media joint representation with sparse and semisupervised regularization, *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6), 2014, 965–978.
- [13] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, *International Journal of Computer Vision*, 106(2), 2014, 210–233.
- [14] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, Semi-supervised cross-media feature learning with unified patch graph regularization, *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3), 2016, 583–596.



- [15] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, RGB-D salient object detection via 3D convolutional neural networks, *Proc. AAAI Conf. on Artificial Intelligence, Virtual Event*, 2021, 1063–1071.
- [16] R. Sun, L. Xuan, L. Hongyan, and W. Lu, Cultivated land segmentation of remote sensing image based on PSPNet of attention mechanism, *International Journal of Robotics & Automation*, 37(1), 2022, 11–19.
- [17] G. Bao, and Y. Zhang, Contextualized rewriting for text summarization, *Proc. AAAI Conf. on Artificial Intelligence, Virtual Event*, 2021, 12544–12553.
- [18] D. Wu, P. Liu, and Y. Zou, A novel method for extracting text from a geometric region, *International Journal of Robotics & Automation*, 36(5), 2021, 325–336.
- [19] Y. Peng, J. Qi, X. Huang, and Y. Yuan, CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network, *IEEE Transactions on Multimedia*, 20(2), 2018, 405–420.
- [20] F. Feng, X. Wang, and R. Li, Cross-modal retrieval with correspondence autoencoder, *Proc. ACM Conf. on Multimedia, Univ Cent Florida, Orlando, FL*, 2014, 7–16.
- [21] Y. Peng, X. Huang, and J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, *Proc. 25th International Joint Conf. on Artificial Intelligence*, New York, NY, USA, 2016, 3846–3853.
- [22] J. Yu, H. Zhou, Y. Zhan, and D. Tao, Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing, *Proc. AAAI Conf. on Artificial Intelligence, Virtual Event*, 35(5), 2021, 4626–4634.
- [23] G. Ding, Y. Guo, and J. Zhou, Collective matrix factorization hashing for multimodal data, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, 2083–2090.
- [24] M. Long, Y. Cao, J. Wang, and P.S. Yu, Composite correlation quantization for efficient multimodal retrieval, *Proc. 39th International ACM Sigir Conf. on Research and Development in Information Retrieval*, Pisa, Italy, 2016, 579–588.
- [25] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval, *IEEE Transactions on Multimedia*, 14(8), 2021, 466–479.
- [26] Z. Yang, X. Deng, L. Guo, and J. Long, Asymmetric supervised fusion-oriented hashing for cross-modal retrieval, *IEEE Transactions on Cybernetics*, early access, 2023, 1–14.
- [27] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *Proc. International Conf. on Learning Representations*, San Diego, CA, USA, 2015, 1–15.
- [28] X. Song, F. Feng, X. Han, X. Yang, and F. Feng, Neural compatibility modeling with attentive knowledge distillation, *Proc. ACM/Sigir Proc. 2018*, Univ Michigan, Ann Arbor, MI, 2018, 5–14.
- [29] L. Jing, E. Vahdani, J. Tan, and Y. Tian, Cross-modal center loss for 3D cross-modal retrieval, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Virtual Event*, 2021, 3142–3151.



*Baohua Qiang* received the B.S. and M.S. degrees from Southwest University in 1996 and 2002, respectively, and the Ph.D. degree from Chongqing University in 2005. In 2007, he joined with the University of Illinois as a Visiting Scholar. From 2007 to 2009, he was a postdoctor with the South China University of Technology. He is currently a Professor with the Guilin University of Electronic Technology. His major research interests include web information processing, intelligent search, massive data processing, and network information integration.



*Mingliang Zhou* received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2017. He held a postdoctoral position with the Department of Computer Science, City University of Hong Kong, Hong Kong, from September 2017 to September 2019. He is currently a Lecturer with the School of Computer Science, Chongqing University, Chongqing, China.

His research interests include image and video coding, perceptual image processing, multimedia signal processing, rate control, multimedia communication, as well as machine learning and optimisation.



*Shihao Zhang* is currently pursuing the Ph.D. degree in cyberspace security with the Guilin University of Electronic Technology. His major research interests include computer vision and pose estimation.



*Hong Zheng* received the B.E. and M.A. degrees from the Guilin University of Electronic Technology in 1997 and 2006, respectively, and the Ph.D. degree from Xiamen University in 2018. She is currently a Lecturer with the Guilin University of Electronic Technology. Her major research interests are medical image reconstruction, image processing, and machine learning.

## Biographies



*Ruidong Chen* received the M.S. degree from the Guilin University of Electronic Technology in 2021, where he is currently pursuing the Ph.D. degree. His major research interests include machine learning, deep learning, and cross-modal analysis and retrieval.



*Chenghua Tang* , received the Ph.D. degree in computer science from the Beijing Institute of Technology, China, in 2007. He is currently a Professor with the School of Computer Science and Information Security, Guilin University of Electronic Technology. His main research interests include information security, software code audit, and security policy analysis.