A Web Knowledge-Driven Multi-Modal Retrieval Method in Computational Social Systems: Unsupervised and Robust Graph Convolutional Hashing

Youxiang Duan, Ning Chen, *Graduate Student Member, IEEE,* Ali Kashif Bashir, *Senior Member, IEEE,* Mohammad Dahman Alshehri, *Senior Member, IEEE,* Lei Liu, *Member, IEEE,* Peiying Zhang, *Member, IEEE,* and Keping Yu, *Member, IEEE*

***Abstract*—Multi-modal retrieval has received widespread consideration since it can commendably provide massive related data support for the development of Computational Social Systems (CSS). However, the existing works still face the following challenges: (1) Rely on the tedious manual marking process when extended to CSS, which not only introduces subjective errors but also consumes abundant time and labor costs; (2) Only using strongly aligned data for training, lacks concern for the adjacency information, which makes the poor robustness and semantic heterogeneity gap difficult to be effectively fit; (3) Mapping features into real-valued forms, which leads to the characteristics of high storage and low retrieval efficiency. To address these issues in turn, we have designed a multi-modal retrieval framework based on web knowledge-driven, called *Unsupervised and Robust Graph Convolutional Hashing* (URGCH). The specific implementations are as follows: First, a "*secondary semantic self-fusion*" approach is proposed, which mainly extracts semantic-rich features through pre-trained neural networks, constructs the joint semantic matrix through semantic fusion, and eliminates the process of manual marking; Second, a "*adaptive computing*" approach is designed to construct enhanced semantic graph features through the knowledge-infused of neighborhoods and employs Graph Convolutional Networks for knowledge-fusion coding, which enables URGCH to sufficiently fit the semantic modality gap while obtaining satisfactory robustness features; Third, combined with hash learning, the multi-modality data is mapped into the form of binary code, which reduces storage requirements and improves retrieval efficiency. Eventually, we perform plentiful experiments on the web dataset. The results evidence that URGCH exceeds other baselines about 1%-3.7% in MAPs, displays superior performance in all aspects, and can meaningfully provide multi-modal data retrieval services to CSS.**

***Index Terms*—Computational Social Systems (CSS), Knowledge-infused, Knowledge-fusion, Multi-modal Retrieval, Graph convolutional networks (GCNs), Unsupervised Hashing**

This work is partially supported by the National Natural Science Foundation of China under Grant 62001357, in part by the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-006, in part by the Open Foundation of State Key Laboratory of Integrated Services Networks (Xidian University) under Grant ISN23-09, in part by the Taif University Researchers Supporting Project, Taif University, Taif, Saudi Arabia under Grant TURSP-2020/126, in part by the Guangdong Basic and Applied Basic Research Foundation under 2020A1515110079, in part by the Shandong Provincial Natural Science Foundation, China under Grant ZR2020MF006, in part by the China Postdoctoral Science Foundation under Grant 2021M692501, in part by the Fundamental Research Funds for the Central Universities under Grant 20CX05017A and XJS210107, and in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under grant JP18K18044 and JP21K17736. (*Corresponding author: **Keping Yu***)

Youxiang Duan and Ning Chen are **co-first authors** of this work.

Youxiang Duan, Ning Chen and Peiying Zhang are with the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China. (*e-mail: yxduan@upc.edu.cn; nchen.bupt@gmail.com; zhangpeiying@upc.edu.cn*).

Ali Kashif Bashir is with the Department of Computing and Mathematics, E-154, John Dolton, Chester Street, M15 6H, Manchester Metropolitan University, Manchesterm UK (*e-mail: dr.alikashif.b@ieee.org*).

Mohammad Dahman Alshehri is with the Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia (*e-mail: alshehri@tu.edu.sa*).

Lei Liu is with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China and also with the Xidian Guangzhou Institute of Technology, Guangzhou 510555, China (*e-mail: tianjiaoliulei@163.com*).

Keping Yu is with the Graduate School of Science and Engineering, Hosei University, Tokyo 184-8584, Japan. (*e-mail: keping.yu@ieee.org*).

I. INTRODUCTION

In the era of big data, Computational Social Systems (CSS) [1] has been pushed to the focus of research due to the rapid development of various technologies such as network information systems and the Internet of Things [2]–[5]. The popularity of various related applications has brought large-scale data, which has brought unprecedented challenges to the analysis of social behaviors with complex correlations [6]. Computational social science, an academic sub-discipline, is emerged as the times require [7]. Its main purpose is to use knowledge-driven computers to model, compute, and analyze social (web) data. Many researchers are also working to discover the social phenomena hidden in the increasingly complex large-scale social data, such as social network analysis [8], COVID-19 analysis [9], public opinion analysis [10], sentiment analysis [11], social media content analysis [12], similarity analysis [13], etc. They analyze social behaviors on multiple dimensions and levels to promote the further development of CSS. The development of these related technologies often requires the driving support of interrelated multi-modal data. More importantly, how to retrieve more modality supplementary data through one modality data to support these technologies is a key challenge to be solved [14]. Therefore, in this research, a multi-modal retrieval method is designed based on social knowledge-driven in CSS.

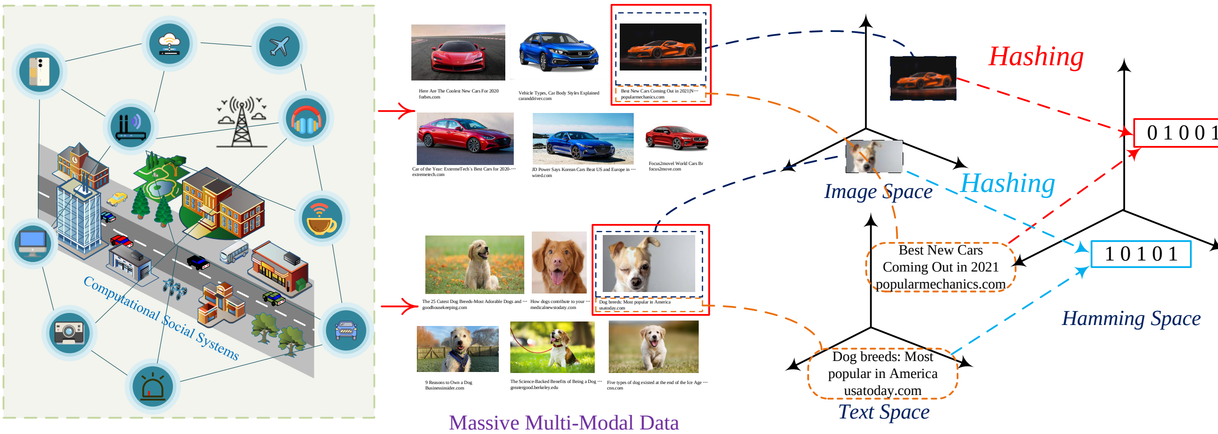


Fig. 1. Multi-modal retrieval in social computing system

As shown in Fig. 1, since the modality data in different feature spaces have separate distribution structures, they fail to be directly compared. Therefore, we ought to map them to the same common subspace for similarity comparison while preserving the original semantic similarity [13]. In the scenario of enormous social datasets, low storage and real-time retrieval have brought tremendous challenges [15], [16]. The hash learning effectively reduces storage requirements and improves retrieval efficiency by mapping original data as compact binary codes in common subspace, *i.e.*, Hamming space. It is worth noting that similar instances in the common subspace should have the same similarity to the original space, *i.e.*, the principle of *preserving similarity* [17].

Traditional multi-modal retrieval methods all utilize artificially annotated semantic labels for supervised training. However, for the CSS, this will bring massive time and labor costs [18], so such supervised learning methods will significantly reduce the generalization of the model [19]. Therefore, we recommend using unsupervised learning to solve this difficulty. Different from the previous methods to obtain the semantic matrix in the process of complicated marking, we utilize the "*secondary semantic self-fusion*" to automatically construct it, which considerably saves time and labor costs.

For data co-occurring in the network, due to human subjectivity and other reasons, its relevance is often weak or irrelevant [20]. Consequently, in the CSS, there are massive irregular, disordered, and unstructured data [21], [22], meantime in what way to enhance the robustness to actual data and avoiding the prediction errors caused by various factors is a challenging issue. Traditional works only exploit strong alignment data to train, which lacks consideration of this problem and makes it often unable the expected effect to deal with actual data [15]. How to conduct training in the face of data that is weak relevant even irrelevant? In this work, we utilize this kind of enhanced sample to train by constructing graph features through the knowledge-infused of the adjacency relationship between semantically related instances. A feature encoder based on the Graph Convolutional Network (GCNs) [23] is designed, which combines with hash learning for knowledge-fusion by employing the semantics of adjacent points, thereby enhancing the robustness of the model.

To sum up, in this research, the *contributions* can be outlined:

- In the real scene of CSS, to meaningfully provide multi-modal data retrieval services, an innovative unsupervised multi-modal retrieval method is proposed, called Unsupervised and Robust Graph Convolutional Hashing (URGCH), which is an end-to-end framework based on social knowledge-driven and primarily comprises two parts: "*Knowledge-Infused*" and "*Knowledge-fusion*".
- To address the issue of the tedious manual marking process and multitudinous time costs, we proffer a "*secondary semantic self-fusion*" method to automatically construct the joint semantic matrix, which is used to bridge the modality gap and guide the training process of URGCH.
- To address issues of poor robustness, fit of the heterogeneous gap, and the requirements of low storage and high retrieval efficiency, a method of "*adaptive computing*" is proposed, which constructs enhanced semantic graph features based on the knowledge-infused of the adjacency relationship between semantically related instances. At the same time, we employ GCNs to perform hash mapping and update its features with the semantic information of adjacent points by knowledge-fusion to enhance the robustness.
- Finally, plentiful experiments have been performed, and the results manifest that URGCH surpasses other baselines to show more satisfactory performance. The specific conditions of each metric are as follows. *Mean average precision* (MAP) has improved by 1%-3.7%, *topK-precision* has surpassed other baselines, the *actual retrieval results* are also satisfactory, and the framework quickly converges about $6-7$ iterations during the training process. Conclusively, URGCH driven by social knowledge can meaningfully provide multi-modal retrieval services for CSS.

The remaining of this article is arranged as follows: In Section II, related works have been analyzed and stated. Subsequently, the problem definition and proposed URGCH are presented in detail in Section III. In Section IV, the corresponding experimental procedures, results, and analysis have been carefully provided. Finally, the work done is condensed in Section V.

II. Related Work

The social knowledge-driven multi-modal retrieval methods can preferably provide data support for various kinds of research on CSS. Therefore, unsupervised multi-modal retrieval has gradually attracted widespread attention in the academic community. Under the premise of not using semantic labels, it mainly preserves the similarity of heterogeneous data through co-occurrence information between modality data. According to the different ways of feature extraction, it can be split into the methods of shallow structure-based and deep structure-based.

Shallow structure-based: As one of the earliest unsupervised cross-modal hashing methods, Inter-Media Hashing (IMH) [24] extends spectral hashing [25] to the field of multi-modal. It explores the similarity between modality data by calculating the modality similarity in the Hamming space. Based on this work, Collective Matrix Factorization Hashing (CMFH) [26] is the first to utilize matrix factorization technology to fit the modality hash functions, and bridge the modality gap by merging multiple information sources. To value the intrinsic structural representation of features, with the aid of the Hadamard matric, Latent Structure Discrete Hashing Factorization (LSDHF) [27] decomposes similar structures in an unsupervised manner to further strengthen modality associations. However, this kind of shallow structure method is difficult to fully explore the semantic information of modality data through an independent manual feature encoding process [28], which reduces the effectiveness of hash encoding [13], [19].

Deep structure-based: Due to its rich nonlinear representation ability [6], [22], [29], the extracted features of deep networks contain richer semantic information and are more discriminative and effective [15] in multi-modal retrieval. Unsupervised Deep Cross-Modal Hashing (UDCMH) [30] combines deep learning, matrix factorization technology [31], and binary latent factor models [32] to jointly optimize feature learning and hash code learning. In addition, it does not need to relax and directly generate unified hash codes. To enable the learned hash codes to maintain the neighborhood structure of the original modality data, Deep Joint-Semantics Reconstructing Hashing (DJSRH) [33] constructs a novel joint semantic matrix to capture latent semantic affinity. And in the training process, the aforementioned matrix is reconstructed to the greatest extent, consequently, better performance is obtained. To fully and effectively capture the correlation between modality data and enhance the discriminative ability of hash codes, Joint-modal Distribution-based Similarity Hashing (JDSH) [28] constructs a joint matrix to preserve semantic similarity, meanwhile using a method based on sampling and weighting to generate hash codes of more discriminative. To provide reliable guidance to further fit cross-modal differences, Aggregation-based Graph Convolutional Hashing (AGCH) [34] proposes a more efficient retrieval strategy. Specifically, without semantic supervision, it uses a variety of similarity measures to measure the structural information of modalities from multiple perspectives, and finally obtains a similarity matrix through the aggregation strategy. Reconstruction Regularized Low-rank Subspace Learning (RRLSL) [35] recovers modality information through the latent representation of optimal conditions, which can effectively deal with scenarios with missing semantic labels. JOint-teachingG (JOG) [36] provides a lightweight and high-performance unsupervised cross-modal retrieval framework, which mainly uses pre-trained models to guide the learning of the trained models. And a refinement strategy is designed to remove random noise, which further improves the model performance through joint learning.

Although these methods exhibit respectable performance, they are difficult to meet expectations when dealing with real data in CSS. In addition, strong alignment data are utilized to explore modality co-occurrence information, which makes the modality semantic information underutilized and the modality gap difficult to fit.

TABLE I
NOTATIONS

Description	Notation
scalar	x
vector	$\boldsymbol{x}$
matrix	$\boldsymbol{X}$
the i-th row of matrix	$\boldsymbol{X}_{i*}$
the j-th column of matrix	$\boldsymbol{X}_{*j}$
the element in i-th row and j-th column of matrix	$\boldsymbol{X}_{ij}$
the transpose of matrix	$\boldsymbol{X}^T$
the Frobenius norm of matrix	$\lVert\boldsymbol{X}\rVert_F$
the trace of matrix	$tr(\boldsymbol{X})$
element sign function	$\text{sign}(x) = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$

TABLE II
DEFINED NOTATION OF THE DATA

Notation	Description
n	the number of the data
k	the length of hash codes
d_v	the dimension of image instance
d_t	the dimension of text instance
$\boldsymbol{V} \in \mathcal{R}^{n \times d_v}$	the original data of image
$\boldsymbol{T} \in \mathcal{R}^{n \times d_t}$	the original data of text
$\boldsymbol{S} \in \mathcal{R}^{n \times n}$	the similarity matrix
$\boldsymbol{D} = \{\boldsymbol{V}_{i*}, \boldsymbol{T}_{i*}\}_{i=1}^{n}$	the training data

III. The Proposed Approach URGCH

In this section, we have introduced the problem definition, configuration information, coding and learning process of the URGCH in detail.

A. Problem Definition

In this research, we concentrate on bimodal multi-modal retrieval, *i.e.*, image and text. Without loss of generality, more modalities can be effortlessly expanded. The relevant notations employed are recorded in Table I. Accordingly, the defined

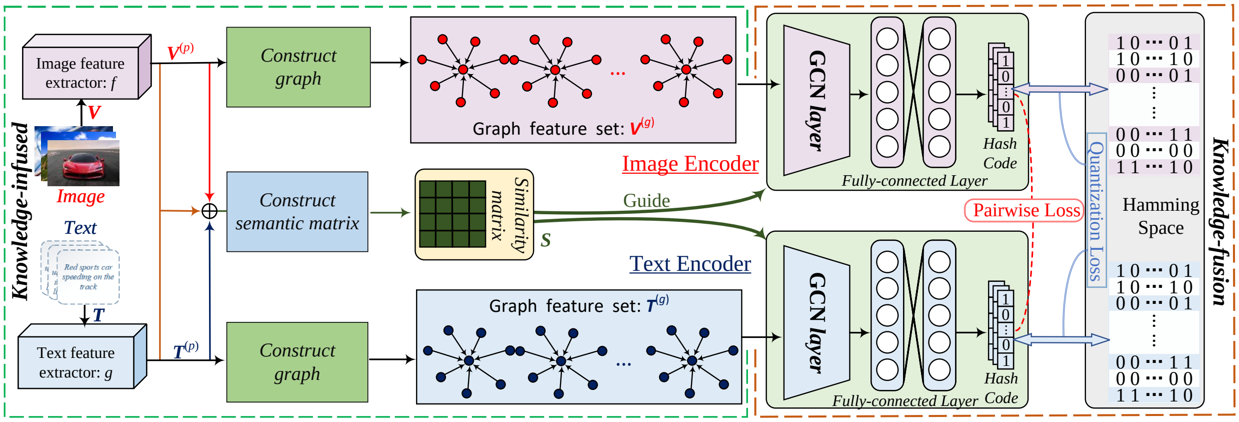


Fig. 2. The framework of URGCH, which includes *Image Encoder* and *Text Encoder*. The encoding process consists of two parts: (a). "*Knowledge-Infused*" and (b). "*Knowledge-fusion*". (a). The original modality data $\boldsymbol{V}$ and $\boldsymbol{T}$ based on the knowledge-infused of the adjacency relationship undergo the "*Construct graph*" process to obtain the graph features $\boldsymbol{V}^{(g)}$ and $\boldsymbol{T}^{(g)}$, respectively, and the "*Construct semantic matrix*" process to obtain the joint semantic matrix $\boldsymbol{S}$. (b). It includes two independent coding parts, which are mainly used to map the semantically rich graph features into the Hamming embedding subspace through the knowledge-fusion.

notations of the data employed in this work are recorded in Table II.

B. Model

The flow chart of URGCH is graphically displayed in Fig. 2, mainly including two parts: "*Knowledge-Infused*" and "*Knowledge-Fusion*", which will be described as follows.

1) Knowledge-Infused: In this subsection, the main purpose is to obtain graph features and joint a semantic matrix for the subsequent "*Knowledge-Fusion*". Therefore, we propose the approach of "*secondary semantic self-fusion*" to construct the joint semantic matrix, and "*adaptive computing*" to construct graph features.

a) Secondary semantic self-fusion to construct semantic matrix: In the supervised multi-modal retrieval methods [15], [37], the semantic matrix is constructed using artificially annotated labels for supervising the training process. However, the marking process in CSS requires massive time and labor costs. In the unsupervised scene, the multi-label annotations are unable to be used, so there is no way to construct the traditional pairwise multi-label semantic matrix. At the same time, the rich semantic similarity implied in the data is essential to bridging the modality gap. The features derived by the deep neural network contain the rich semantics in the original data. Therefore, without using multi-label labels, the semantic features derived by the feature extractor are employed to build the similarity matrix. In this work, the proposed approach of construction is called "*secondary semantic self-fusion*", which is based on the *cosine distance*, and illustrated in Fig. 3.

For the image modality, let the semantic feature obtained after the feature extractor f be $\boldsymbol{V}^{(p)} = \left\{v_{i*}^{(p)}\right\}_{i=1}^{n}$, and the image semantic matrix $\boldsymbol{S}^{(v)}$ is defined as:

$$\boldsymbol{S}^{(v)} = \left\{s_{ij}^{(v)}\right\}^{n\times n} = \left\{\frac{v_{i*}^{(p)}\left(v_{j*}^{(p)}\right)^{T}}{\left\|v_{i*}^{(p)}\right\|_2\left\|v_{j*}^{(p)}\right\|_2}\right\}^{n\times n}, \quad s.t. \quad s_{ij}^{(v)} \in [-1,+1]. \tag{1}$$

For the text modality, let the semantic feature obtained by the feature extractor g be $\boldsymbol{T}^{(p)} = \left\{t_{i*}^{(p)}\right\}_{i=1}^{n}$. It is worth noting that each $t_{i*}^{(p)}$ after feature extraction is still related to $v_{i*}^{(p)}$. The text semantic matrix is formulized as:

$$\boldsymbol{S}^{(t)} = \left\{s_{ij}^{(t)}\right\}^{n\times n} = \left\{\frac{t_{i*}^{(p)}\left(t_{j*}^{(p)}\right)^{T}}{\left\|t_{i*}^{(p)}\right\|_2\left\|t_{j*}^{(p)}\right\|_2}\right\}^{n\times n}, \quad s.t. \quad s_{ij}^{(t)} \in [-1,+1]. \tag{2}$$

To preserve the uniformity of semantic distribution between modalities, we merge the image and text semantic matrix into a unified similarity matrix $\boldsymbol{S}$, which is called "*semantic secondary fusion*". The specific integration method is as follows:

$$\boldsymbol{S} = \{s_{ij}\}^{n\times n} = \lambda\boldsymbol{S}^{(v)} + \zeta\boldsymbol{S}^{(t)} + \xi cos(\boldsymbol{S}^{(v)}, \boldsymbol{S}^{(t)}), \quad s.t. \quad s_{ij} \in [-1,+1],\ \lambda+\zeta+\xi=1, \tag{3}$$

where λ, ζ and ξ are hyperparameters. We use the validation set to adjust adaptively to obtain the best weight distribution, which will be explained in detail in Section IV-G. To maintain the semantic distribution between the modalities, *i.e.*, as shown in Fig. 3 that the instances corresponding to the semantically similar instances in one of the modalities in another modality should also be similar, so we introduce the cosine similarity of the third term. $\boldsymbol{S}$ records the pairwise similarity of the image-text in the dataset, and fully integrates the semantic distribution information between the modalities into a unified joint matrix. Therefore, in the unsupervised scenario of the CSS, we utilize the joint semantic similarity matrix $\boldsymbol{S}$ to guide the semantic *preserving similarity* of the model training process.

b) Adaptive computing to construct graph feature: The original image data $\boldsymbol{V}$ and the text data $\boldsymbol{T}$ pass the feature extractors f and g to obtain feature representation matrices $\boldsymbol{V}^{(p)}$ and $\boldsymbol{T}^{(p)}$ that contain rich semantics, respectively. To input them into the GCN layers, we need to construct graph features. The existing supervised graph construction [17]

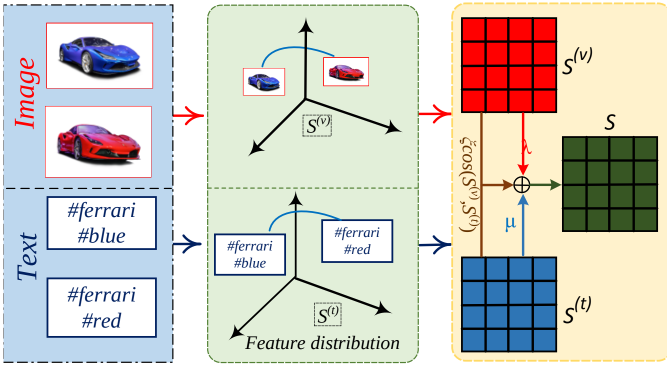


Fig. 3. Secondary Semantic Self-Fusion

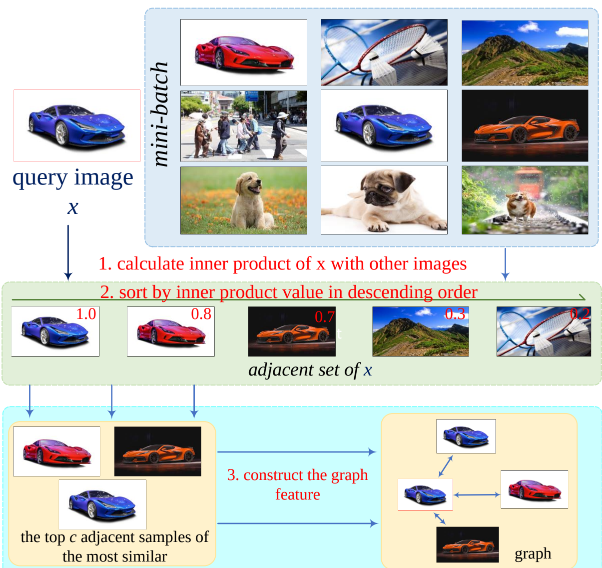


Fig. 4. Adaptive Computing to construct graph feature, taking the image modality as an example.

constructs the adjacency matrix through the artificial multi-label annotation relationship between the data, nevertheless, the multi-label annotation cannot be used in the unsupervised environment. Therefore, we put forward a novel approach to unsupervised graph feature construction, called "*adaptive computing*", which is illustrated in Fig. 4.

Taking image modalities as an example, given the query image x, the purpose of constructing graph features is to find out the data related to it in the mini-batch as much as possible. Then, the graph features are input into the GCN layer to fuse and strengthen the semantic knowledge of samples and improve the robustness of URGCH. As we all know, the greater the similarity between two instances with a larger inner product [15]. Therefore, we calculate the inner product by calculating x and all samples in the mini-batch $\boldsymbol{V}^{(p)}$, and obtain the adjacent set x_{adj} of sample similarity sorted from large to small as follows:

$$x_{adj} = rsort\left(\left\{\langle x, v_{i*}^{(p)}\rangle\right\}_{i=1}^{n}\right), \tag{4}$$

where $rsort()$ is the descending operation, and $\langle\rangle$ is the inner product. At the same time, to eliminate random errors and ensure robustness, we set the top c adjacent points of most similar to construct the graph features $x^{(g)}$, as follows:

$$x^{(g)} = \left\{(x_{adj})_i\right\}_{i=1}^{c}, \tag{5}$$

where we set $c = 10$. Through the above steps, the graph feature $x^{(g)}$ of the image x can be obtained.

Similarly, we can obtain the graph feature set $\boldsymbol{V}^{(g)} = \left\{v_{i*}^{(g)}\right\}_{i=1}^{n}$ of the image data set $\boldsymbol{V}^{(p)}$ and the graph feature set $\boldsymbol{T}^{(g)} = \left\{t_{i*}^{(g)}\right\}_{i=1}^{n}$ of the text data set $\boldsymbol{T}^{(t)}$, respectively.

2) Knowledge-Fusion: In this subsection, the main purpose is to use the joint semantic matrix to guide the hash coding process for knowledge-fusion. As shown in Fig. 2, it mainly contains two encoders based on GCNs, *i.e.*, the *Image Encoder* and *Text Encoder*, which are used to map the constructed graph features to unified hash codes in the Hamming space. The composition of the two encoders and the corresponding configuration are presented as follows.

a) Image Encoder: The image feature extractor f is derived from CNN-F [38] pre-trained on the ImageNet dataset [39], and the first seven layers of parameters are frozen and used to initialize f. At the same time, the input size of f is adjusted to $3 \times 224 \times 224$.

In the CSS, there are enormous irregular, disordered and unstructured data. Therefore, to strengthen the robustness of URGCH to real data, we explore the encoder network based on multi-layer GCN layers. As mentioned above, the image feature $\boldsymbol{V}$ undergoes "knowledge-infused" to obtain the image feature $\boldsymbol{V}^{(g)}$. At the same time, the set of adjacent points of each image data point x is:

$$\mathcal{N}_x = \left\{j | j \in x^{(g)}, 1 \leq j \leq n\right\}, \tag{6}$$

Correspondingly, we can calculate the adjacency matrix $\boldsymbol{A}^{(v)} = \left\{a_{ij}^{(v)}\right\}_{i=1,j=1}^{n \times n}$ of the image modality undirected graph $\boldsymbol{V}^{(g)}$ as follow:

$$a_{ij}^{(v)} = \begin{cases} 1, & j \in \mathcal{N}_i, \\ 0, & j \notin \mathcal{N}_i. \end{cases} \tag{7}$$

Based on the above information and inspired by work [23], the forward inter-layer propagation rules of the multi-layer GCN adopt the following forms:

$$\boldsymbol{H\_v}^{(l+1)} = \sigma\left(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}^{(v)}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H\_v}^{(l)}\boldsymbol{W\_v}^{(l)}\right), \tag{8}$$

where $\widetilde{\boldsymbol{A}}^{(v)} = \boldsymbol{A}^{(v)} + \boldsymbol{I}_n$. $\boldsymbol{I}_n$ is the n-dimensional identity matrix, which means that each node is connected to itself so that the features of the vertex itself are also preserved. $\widetilde{\boldsymbol{D}}$ represents a degree matrix, furthermore, $\widetilde{\boldsymbol{D}}_{ii} = \sum_j \widetilde{\boldsymbol{A}}_{ij}^{(v)}$. $\boldsymbol{H\_v}^{(l)}$ indicates the input features of the l-th GCN layer. Moreover, its dimension of output is F^l and weight matrix is denotes as $\boldsymbol{W\_v}^{(l)}$, which will be continuously learned and updated during the training process. The dimension of $\boldsymbol{W\_v}^{(l)}$ is $F^l \times F^{(l+1)}$. $\boldsymbol{H\_v}^{(l+1)}$ indicates the input features of the (l+*1*)-th GCN layer and the output features of the l-th GCN layer. $\sigma(\cdot)$ represents the nonlinear activation function, where

TABLE III
DIMENSIONAL CONFIGURATION INFORMATION OF *Image Encoder*

Number	Layer	Dimension
1	CNN-F	4096
2	GCN_1	1024
3	GCN_2	512
4	fc_1	512
5	fc_2	k

the *LeakyReLU* activation function is used in the GCN layer. In multi-layer GCN, we construct graph features from the most relevant samples of top c and meanwhile integrate its features and the features of samples with complementary semantics, to sufficiently strengthen the robustness of URGCH in CSS.

On this basis, $fc_1 \rightarrow fc_2$ (fully-connected layers) have been added to map features to hash codes in the joint Hamming space. In conclusion, the dimensional configuration information of *Image Encoder* is recorded in Table III, where the number in the column of "**Dimension**" betokens the output dimension of this layer.

b) Text Encoder: To encode text data, it is first expressed as bag-of-words vectors, which will be input to the text feature extractor g. We construct two fully-connected layers ($fc_1 \rightarrow fc_2$) as text feature extractor, which is mainly used to extract rich semantic features $\boldsymbol{T}^{(p)}$. It is worth noting that the parameters of g will also be optimized and updated during the learning process, which will be thoroughly explained in Section III-C.

Similarly, the encoding process is similar to the image encoder f. After obtaining the graph feature $\boldsymbol{T}^{(g)}$, the set of adjacent points of each text y is:

$$\mathcal{M}_y = \left\{ j | j \in y^{(g)}, 1 \leq j \leq n \right\}, \tag{9}$$

The adjacency matrix $\boldsymbol{A}^{(t)} = \left\{ a_{ij}^{(t)} \right\}_{i=1,j=1}^{n \times n}$ of the text modality undirected graph $\boldsymbol{T}^{(g)}$ is as follows:

$$a_{ij}^{(t)} = \begin{cases} 1, & j \in \mathcal{M}_i, \\ 0, & j \notin \mathcal{M}_i. \end{cases} \tag{10}$$

Correspondingly, the forward inter-layer propagation rules of the multi-layer GCN is expressed in Eq. 11:

$$\boldsymbol{H\_t}^{(l+1)} = \sigma \left(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}} \widetilde{\boldsymbol{A}}^{(t)} \widetilde{\boldsymbol{D}}^{-\frac{1}{2}} \boldsymbol{H\_t}^{(l)} \boldsymbol{W\_t}^{(l)} \right), \tag{11}$$

In the same way, we add two fully-connected layers ($fc_3 \rightarrow fc_4$) after the GCN layers for hash mapping. Eventually, the dimensional configuration information of *Text Encoder* is demonstrated in Table IV.

It must be noted that the *Image Encoder* and the *Text Encoder* are independent networks. In this article, although some parameters are the same in form, their contents are independent of each other and are not shared.

C. Learning

Let $\boldsymbol{F} = Img\_Encoder\left(\boldsymbol{V}; \theta^{(v)}\right)$ represent the final output features of *Image Encoder*, where $\theta^{(v)}$ denotes its parameters.

TABLE IV
DIMENSIONAL CONFIGURATION INFORMATION OF *Text Encoder*

Number	Layer	Dimension
1	fc_1	4096
2	fc_2	4096
3	GCN_1	1024
4	GCN_2	512
5	fc_3	512
6	fc_4	k

Similarly, let $\boldsymbol{G} = Txt\_Encoder\left(\boldsymbol{T}; \theta^{(t)}\right)$ represent the final output features of the *Text Encoder*, where $\theta^{(t)}$ denotes its parameters. Our purpose is to continuously optimize and update parameters of *Encoders* through the learning process.

Consequently, the objective function of URGCH have been designed to be:

1) Image Encoder:

$$\begin{aligned} \min_{\theta^{(v)}} \mathcal{L}^{(v)} &= \mathcal{L}_1^{(v)} + \alpha \mathcal{L}_2^{(v)} \\ &= - \sum_{i,j=1}^{n} \left(\boldsymbol{S}_{ij} \Delta_{ij}^{(v)} - \log\left(1 + e^{\Delta_{ij}^{(v)}}\right) \right) \\ &\quad + \alpha \left\| \boldsymbol{F} - \boldsymbol{B}^{(v)} \right\|_F^2, \end{aligned} \tag{12}$$

where $\Delta_{ij}^{(v)} = \frac{1}{2} \boldsymbol{F}_{i*} \boldsymbol{G}_{j*}^T$, $\boldsymbol{B}^{(v)} = \text{sign}(\boldsymbol{F}) \in \{-1, +1\}^{n \times k}$ is the predicted hash codes of *Image Encoder*, α is the hyperparameter. $\mathcal{L}_1^{(v)}$ is the negative log-likelihood, and optimizing this term is equivalent to the maximum likelihood. By optimizing this term, the semantic consistency and relevance between the original data can be well preserved, which can be derived from the following formula:

$$p\left(S_{ij} \mid \mathbf{F}_{i*}, \mathbf{G}_{j*}\right) = \begin{cases} \sigma\left(\Delta_{ij}\right), & S_{ij} = 1, \\ 1 - \sigma\left(\Delta_{ij}\right), & S_{ij} = 0, \end{cases} \tag{13}$$

where $\sigma(\Delta_{ij}) = \frac{1}{1+e^{-\Delta_{ij}}}$. Δ represents a measure of similarity in the form of inner product. Therefore, the more similar between $\boldsymbol{F}_{i*}$ and $\boldsymbol{G}_{j*}$, the greater the inner product and the higher the probability. $\mathcal{L}_2^{(v)}$ is the quantization loss, which is utilized to minimize the mistake of learning the hash codes.

2) Text Encoder:

$$\begin{aligned} \min_{\theta^{(t)}} \mathcal{L}^{(t)} &= \mathcal{L}_1^{(t)} + \alpha \mathcal{L}_2^{(t)} \\ &= - \sum_{i,j=1}^{n} \left(\boldsymbol{S}_{ij} \Delta_{ij}^{(t)} - \log\left(1 + e^{\Delta_{ij}^{(t)}}\right) \right) \\ &\quad + \alpha \left\| \boldsymbol{T} - \boldsymbol{B}^{(t)} \right\|_F^2, \end{aligned} \tag{14}$$

where $\Delta_{ij}^{(v)} = \frac{1}{2} \boldsymbol{G}_{i*} \boldsymbol{F}_{j*}^T$, the predicted hash codes is represented as $\boldsymbol{B}^{(t)} = \text{sign}(\boldsymbol{G}) \in \{-1, +1\}^{n \times k}$. $\mathcal{L}_1^{(t)}$ and $\mathcal{L}_2^{(t)}$ are similar losses as in *Image Encoder*.

As a consequence, combining Eq. 12 and Eq. 14, the overall objective function can be represented as:

$$\min_{\theta^{(v)}, \theta^{(t)}} \mathcal{L} = \mathcal{L}^{(v)} + \mathcal{L}^{(t)}. \tag{15}$$

Algorithm 1: The learning algorithm of URGCH

Input: $\boldsymbol{V}$; $\boldsymbol{T}$; k.
Output: $\theta^{(v)}$; $\theta^{(t)}$; $\boldsymbol{B}^{(v)}$ and $\boldsymbol{B}^{(t)}$ (hash codes).
Initialization: α, λ, ζ and ξ; $\theta^{(v)}$ and $\theta^{(t)}$; μ (*learning rate*); $m = 128$ (*batch size*); $t = \lceil \frac{n}{m} \rceil$ (*number of iterations*); $e = 200$ (*number of cycle epochs*) and $iter$ (*current iteration*).
repeat
  for $iter = 1, 2, \cdots, t$ **do**
    $\star$ *Randomly take out* m instances from $\boldsymbol{D} = \{\boldsymbol{V}, \boldsymbol{T}\}$ to fabricate a mini-batch;
    $\star$ *Obtain* the semantic features $\boldsymbol{V}^{(p)}$ and $\boldsymbol{T}^{(p)}$ which are extracted by the feature extractor respectively;
    $\star$ *Calculate* the similarity matrix $\boldsymbol{S}$ as reported by Eq.3;
    $\star$ *Construct* the graph feature $\boldsymbol{V}^{(g)}$ and $\boldsymbol{T}^{(g)}$ according to Eq.5;
    $\star$ *Obtain* $\boldsymbol{B}^{(v)}$ and $\boldsymbol{B}^{(t)}$ of through the forward-propagation;
    $\star$ *Calculate* the objective function in Eq.15;
    $\star$ *Use* gradient back-propagation to update parameters $\theta^{(v)}$ and $\theta^{(t)}$.
  end
until *the cycle epoch iterates* e *times*;

In this work, we utilize back-propagation (BP) and mini-batch stochastic gradient descent (SGD) strategies to optimize the objective function $\mathcal{L}$. In Algorithm 1, we summarize the entire workflow of URGCH.

D. Implementation Details

Unified description of the activation function used in URGCH: Unless otherwise specified, the layers that output the predicted hash codes all employ the *tanh*, and the remainder all employ the *ReLU*.

IV. Experiment

In this section, we first introduce the Web social Datasets used in Section IV-A, *MIRFLICKR-25K* and *NUS-WIDE*. Secondly, the Evaluation metrics *Mean average precision* (MAP) and *topK-precision* are introduced in Section IV-B, the baselines used in the comparison experiment are shown in Section IV-C. In addition, it also introduces the related parameters setting not mentioned above in Section IV-D. In Section IV-E, the results of *MAP* and *topK-precision* are shown, which can prove the performance of the model, and then prove the effectiveness of the "*secondary semantic self-fusion*" and "*adaptive computing*" we proposed. In Section IV-F, the experiment of the retrieval results is supplemented to further prove the effectiveness of URGCH. Finally, we conduct hyperparameter analysis experiments to verify the selection of hyperparameters in Section IV-G, and convergence analysis experiments to verify the convergence process of the model in Section IV-H, which proves the effectiveness of the framework combined with hash learning.

TABLE V
Statistics of Dataset Division

Dataset	*MIRFLICKR-25K*	*NUS-WIDE*
Train	10,000	10,500
Test (Query)	2000	2000
Retrieval (Database)	23,000	184,577
Total	25,000	186,577

TABLE VI
Related Parameters Setting

Parameter	Setting
batch_size	128
learning_rate	0.0001 - 0.1
cycle epochs	200
number of adjacent points	10
length of hash code	16, 32, 64, 128
hyperparameter	$\alpha = 0.01$ and $\lambda = 0.3$, $\zeta = 0.3$, $\xi = 0.4$

A. Dataset

1) MIRFLICKR-25K: The *MIRFLICKR-25K* dataset [40] collects 25,000 images and text data obtained from the *FLICKR* website. Each text is represented by a bag-of-words (BOW) vector of 500-dimension.

2) NUS-WIDE: The *NUS-WIDE* dataset [41] collects 269,648 image-text pairs data obtained from various websites, each of which contains 1 to 81 labels. We select a total of 186,577 instances of the 10 most frequent labels as training data. Similarly, each text is represented by a BOW vector of 500-dimension.

It is worth noting that we randomly extract and divide the dataset, which is counted in Table V in detail.

B. Evaluation Metric

Mean average precision (MAP) and topK-precision curve are employed to explore the performance of URGCH. The former derives from averaging of average precision (AP) as follows:

$$\mathrm{AP} = \frac{1}{z} \sum_{i=1}^{z} \frac{t_i}{i}, \tag{16}$$

where z symbolizes the quantity of instances in the database related to the current query, and t_i represents the amount of relevant results within the top i samples. Therefore, the MAP can be calculated as follows:

$$\mathrm{MAP} = \frac{1}{n_q} \sum_{j=1}^{n_q} \mathrm{AP}_j, \tag{17}$$

where n_q denotes the amount of samples inside the query set.

The topK-precision indicates the precision of the model when the number of retrieved samples is different.

C. Baseline

To estimate the effectiveness of URGCH, which has been compared with four state-of-the-art baselines, including

TABLE VII
MAPs results. The best MAPs are shown in boldface.

Retrieval Task	Method	*MIRFLICKR-25K* 16 bits	32 bits	64 bits	128 bits	*NUS-WIDE* 16 bits	32 bits	64 bits	128bits
$Image \rightarrow Text$	CMFH [26]	0.621	0.624	0.625	0.627	0.455	0.459	0.465	0.467
	UDCMH [30]	0.689	0.698	0.714	0.717	0.511	0.519	0.524	0.558
	DJSRH [33]	0.810	0.843	0.862	0.876	0.724	0.773	0.798	0.817
	JDSH [28]	0.832	0.853	0.882	0.892	0.736	0.793	0.832	0.835
	URGCH	**0.859**	**0.871**	**0.901**	**0.914**	**0.773**	**0.820**	**0.842**	**0.859**
	improvement	⇑ 0.027	⇑ 0.018	⇑ 0.019	⇑ 0.022	⇑ 0.037	⇑ 0.027	⇑ 0.010	⇑ 0.024
$Text \rightarrow Image$	CMFH [26]	0.642	0.662	0.676	0.685	0.529	0.577	0.614	0.645
	UDCMH [30]	0.692	0.704	0.718	0.733	0.637	0.653	0.695	0.716
	DJSRH [33]	0.786	0.822	0.835	0.847	0.712	0.744	0.771	0.789
	JDSH [28]	0.825	0.864	0.878	0.880	0.721	0.795	0.794	0.804
	URGCH	**0.853**	**0.888**	**0.895**	**0.907**	**0.758**	**0.809**	**0.822**	**0.836**
	improvement	⇑ 0.028	⇑ 0.024	⇑ 0.017	⇑ 0.027	⇑ 0.037	⇑ 0.014	⇑ 0.028	⇑ 0.032

shallow-structure (CMFH [26]) and deep structure (UDCMH [30], DJSRH [33], JDSH [28]).

To guarantee fairness, all baselines, including the shallow structure, employ the pre-trained CNN-F to extract image features. It is worth noting that the code of UDCMH is not yet open-source, so we implemented it carefully in accordance with the original paper. Moreover, the source codes of other baselines are graciously offered by the original authors, and the corresponding configuration is strictly implemented following the original paper. To ensure interference from other factors, we adopt the unified dataset after the above processing for comparative experiments.

D. Related Parameters Setting

It should be noted that owing to image extractor f using pre-trained CNN-F, we freeze its parameters so that they will not be updated during the learning process. Besides, all parameters of URGCH are initialized randomly and continuously optimized and updated during learning. In the experiment, we set the batch size to 128 and training iteration to 200 times. Furthermore, the learning rate adaptively adjusts from 0.0001 to 0.1 by using the validation set. At the same time, we conduct ten experiments and average the results to eliminate randomness. Finally, we record the relevant parameters and their setting used in this work in Table VI.

E. Performance

1) MAP: Table VII records the MAPs value results of URGCH and all baselines, where "$Image \rightarrow Text$" indicates using the image (query) to retrieve text (database), and "$Text \rightarrow Image$" indicates using the text (query) to retrieve image (database). It can be inferred from the comparison of MAPs values that USGCH can effectively achieve better performance than other baselines. By way of illustration, on *NUS-WIDE* while "$Image \rightarrow Text$" and the length of hash codes is 16 bits, URGCH improves the MAP value by 0.037 compared to the second-best method (JDRH).

2) topK-precision: It has been presented in Fig. 5, including URGCH and all baselines, where the length of hash codes is 128 bits. As is well-known, the higher of the curve, the stronger of performance. Therefore, it can be found that

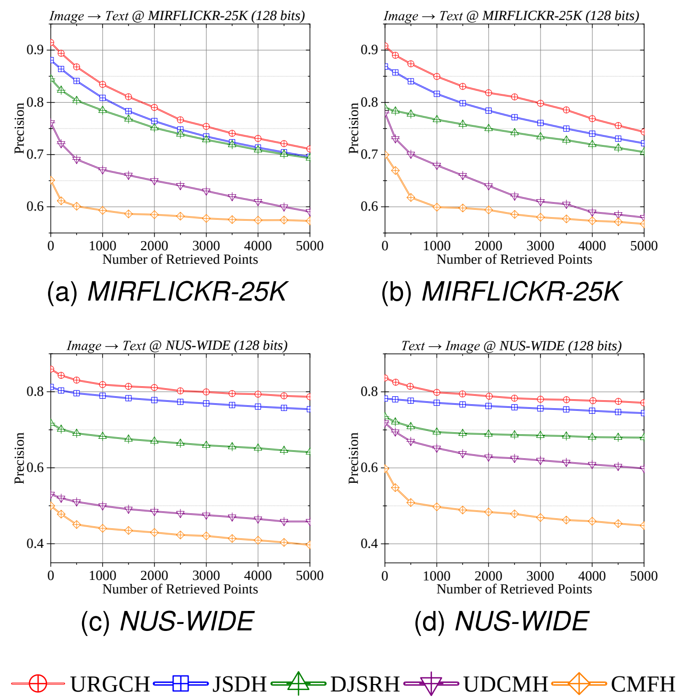


Fig. 5. The topK-precision curves.

URGCH achieves satisfactory performance and outperforms other baselines.

F. Retrieval Results

To verify the actual action of URGCH, two samples (data point of *8916-th*) are tried on MIRFLICKR with the hash codes is set to 32 bits. The results are displayed in Fig. 6, where the left column represents the query, and the right column represents the retrieved results. In addition, "$Text \rightarrow Image$" means using texts as the query to retrieve the image database, and "$Image \rightarrow Text$" means using images as the query to retrieve the text database. Finally, the Hamming distance between the query and the samples in the database is calculated and sorted. Therefore, URGCH can achieve satisfactory multi-modal retrieval tasks.

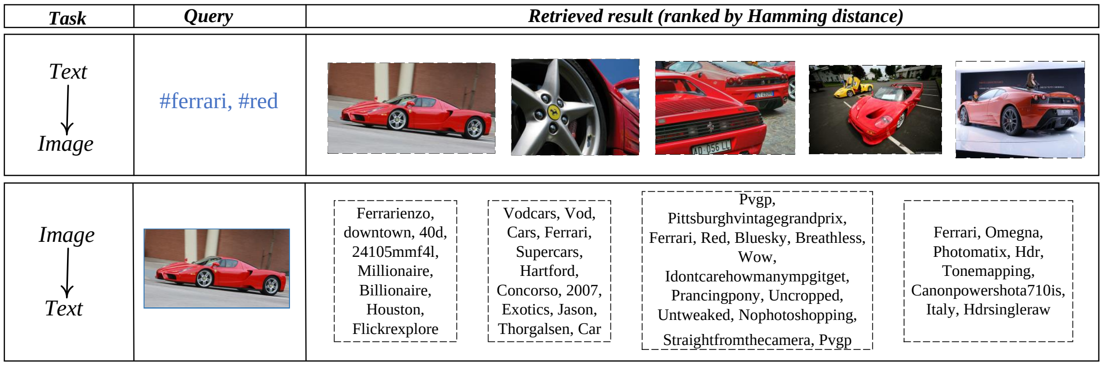
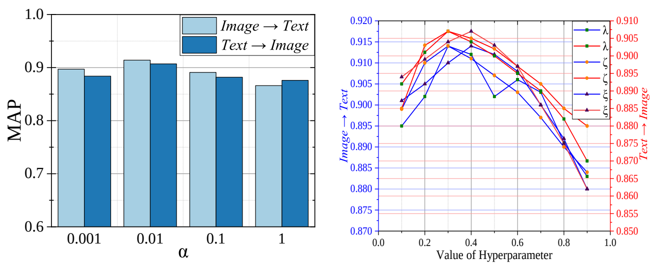


Fig. 6. Retrieved result on *MIRFLICKR-25K*



(a) α (b) λ, ζ and ξ

Fig. 7. The influences of hyperparameters: (a) α in objective function. (b) λ, ζ and ξ in "Secondary semantic self-fusion"

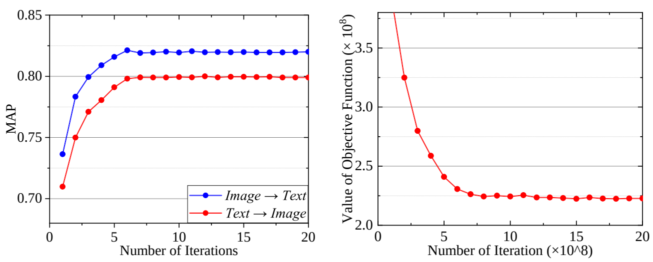


Fig. 8. Convergence curve on *NUS-WIDE*

G. Hyperparameter Analysis

To research the affect of the hyperparameter α in Eq. 15 and the hyperparameters λ, ζ and ξ in the Section III-B1a, we randomly extract 2000 points from *MIRFLICKR-25K* as the validation set for experimentation, where the hash codes is set to 128 bits. The influence of the MAP on different hyperparameters is shown in Fig. 7. Consequently, it can be inferred that the USGCH can achieve the best performance when $\alpha = 0.01$ and $\lambda = 0.3$, $\zeta = 0.3$, $\xi = 0.4$.

H. Convergence Analysis

We conduct an experiment in *NUS-WIDE* to verify the convergence of URGCH, where the length of hash codes is 32 bits. Fig. 8 manifests the variation of the value of objective function and MAP along with the iteration. it has been inferred that the MAP gradually increases as the objective function decreases during the training process, and finally converges quickly about 6-7 iterations.

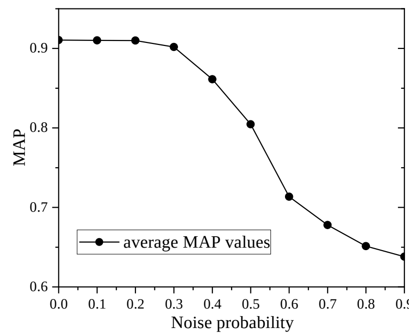


Fig. 9. Robustness analysis on $MIRFLICKR-25K$.

I. Robustness Analysis

Supervised methods usually construct noisy data by randomly changing semantic labels, etc., and then evaluate the robustness of the model. However, semantic labels are not available in unsupervised methods. Therefore, inspired by this, we randomly permutate some element values in the unsupervised semantic matrix constructed in Section III-B1a according to different probability values to introduce some noise and use the trained URGCH for testing. On the $MIRFLICKR-25K$ dataset, we conduct an experimental evaluation of robustness and record the average MAP values for two retrieval tasks "$Image \rightarrow Text$" and "$Text \rightarrow Image$", as shown in Fig. 9, where the length of the hash code is 128 bits. It can be found that when the random noise probability is within 0.3, the performance of URGCH is not significantly affected, and its performance is still excellent. Therefore, URGCH has outstanding robustness. However, when the random noise probability is greater than 0.3, the excessive semantic relations are severely disrupted, so the performance starts to decline significantly.

V. Conclusion and Outlook

In this work, to provide reliable multi-modal retrieval services for CSS, we propose the Unsupervised and Robust Graph Convolutional Hashing (URGCH). It utilizes "*secondary semantic self-fusion*" to construct the joint semantic matrix which is employed to guide the training process, saving abundant time and labor costs in the process of manual marking. Moreover, through the knowledge-infused of the neighborhood, the semantic-enhanced graph features are constructed through the approach of "*adaptive computing*", and the multi-layer GCNs layers are designed for hash coding, which combines with hash learning for knowledge-fusion by employing the semantics of adjacent points and enhances the robustness of URGCH. Finally, extensive experiments on the social dataset demonstrate that URGCH has satisfactory superior performance and can provide multi-modal data support for CSS.

Based on the existing foundation, our future work will be expected to make efforts and breakthroughs in the following points, and expect to dedicate our modest efforts to the development of CSS. (1) Model large-scale noisy datasets in reality to better deal with various real-world scenarios. (2) We found that different adjacencies have primary and secondary importance when constructing graph features. Therefore, it is desirable to apply attention weight to adjacent points to further improve the robustness. (3) Finally, we expect to make further explorations in the extension of modalities, such as audio, video, etc.

References

[1] H. B. Liaqat, A. Ali, J. Qadir, A. K. Bashir, M. Bilal, and F. Majeed, "Socially-aware congestion control in ad-hoc networks: Current status and the way forward," *Future Generation Computer Systems*, vol. 97, pp. 634–660, 2019.

[2] Y. Duan, N. Chen, S. Shen, P. Zhang, Y. Qu, and S. Yu, "Fdsa-stg: Fully dynamic self-attention spatio-temporal graph networks for intelligent traffic flow prediction," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2022, doi: 10.1109/TVT.2022.3178094.

[3] K. Yu, L. Tan, S. Mumtaz, S. Al-Rubaye, A. Al-Dulaimi, A. K. Bashir, and F. A. Khan, "Securing critical infrastructures: Deep-learning-based threat detection in iiot," *IEEE Communications Magazine*, vol. 59, no. 10, pp. 76–82, 2021.

[4] H. Zhang, J. Yu, M. S. Obaidat, P. Vijayakumar, L. Ge, J. Lin, J. Fan, and R. Hao, "Secure edge-aided computations for social internet-of-things systems," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 1, pp. 76–87, 2022.

[5] N. Chen, P. Zhang, N. Kumar, C.-H. Hsu, L. Abualigah, and H. Zhu, "Spectral graph theory-based virtual network embedding for vehicular fog computing: A deep reinforcement learning architecture," *Knowledge-Based Systems*, vol. 257, p. 109931, 2022.

[6] Y. Duan, N. Chen, L. Chang, Y. Ni, S. V. N. S. Kumar, and P. Zhang, "Capso: Chaos adaptive particle swarm optimization algorithm," *IEEE Access*, vol. 10, pp. 29 393–29 405, 2022.

[7] H. Jiang, L. Li, H. Xian, Y. Hu, H. Huang, and J. Wang, "Crowd flow prediction for social internet-of-things systems based on the mobile network big data," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 1, pp. 267–278, 2022.

[8] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, "Sentiment analysis of lockdown in india during covid-19: A case study on twitter," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 992–1002, 2021.

[9] L. Tan, K. Yu, N. Shi, C. Yang, W. Wei, and H. Lu, "Towards secure and privacy-preserving data sharing for covid-19 medical records: A blockchain-empowered approach," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 1, pp. 271–281, 2022.

[10] J. Zhou, H. Zogan, S. Yang, S. Jameel, G. Xu, and F. Chen, "Detecting community depression dynamics due to covid-19 pandemic in australia," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 982–991, 2021.

[11] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, 2020.

[12] P. K. K. N. and M. L. Gavrilova, "Latent personality traits assessment from social network activity using contextual language embedding," *IEEE Transactions on Computational Social Systems*, pp. 1–12, 2021.

[13] P. Zhang, X. Huang, M. Li, and Y. Xue, "Hybridization between neural computing and nature-inspired algorithms for a sentence similarity model based on the attention mechanism," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–21, 2021.

[14] K. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-enhanced data sharing with traceable and direct revocation in iiot," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 11, pp. 7669–7678, 2021.

[15] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3232–3240.

[16] P. Zhang, N. Chen, S. Li, K.-K. R. Choo, and C. Jiang, "Multi-domain virtual network embedding algorithm based on horizontal federated learning," *arXiv preprint arXiv:2205.14665*, 2022.

[17] Y. Duan, N. Chen, P. Zhang, N. Kumar, L. Chang, and W. Wen, "Ms2gah: Multi-label semantic supervised graph attention hashing for robust cross-modal retrieval," *Pattern Recognition*, vol. 128, p. 108676, 2022.

[18] L. Yang, X. Wang, J. Zhang, J. Yang, Y. Xu, J. Hou, K. Xin, and F.-Y. Wang, "Hackgan: Harmonious cross-network mapping using cyclegan with wasserstein-procrustes learning for unsupervised network alignment," *IEEE Transactions on Computational Social Systems*, pp. 1–14, 2022.

[19] K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi, and T. Sato, "Deep-learning-empowered breast cancer auxiliary diagnosis for 5gb remote e-health," *IEEE Wireless Communications*, vol. 28, no. 3, pp. 54–61, 2021.

[20] P. Zhang, X. Huang, and L. Zhang, "Information mining and similarity computation for semi- / un-structured sentences from the social data," *Digital Communications and Networks*, 2020.

[21] B. Guidi, A. Michienzi, and L. Ricci, "A graph-based socioeconomic analysis of steemit," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 2, pp. 365–376, 2021.

[22] Y. Zhang, C. Li, N. Chen, and P. Zhang, "Intelligent requests orchestration for microservice management based on blockchain in software defined networking: a security guarantee," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2022, pp. 254–259.

[23] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[24] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 785–796.

[25] Y. Weiss, A. Torralba, R. Fergus *et al.*, "Spectral hashing." in *Nips*, vol. 1, no. 2. Citeseer, 2008, p. 4.

[26] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5427–5440, 2016.

[27] Y. Fang, B. Li, X. Li, and Y. Ren, "Unsupervised cross-modal similarity via latent structure discrete hashing factorization," *Knowledge-Based Systems*, vol. 218, p. 106857, 2021.

[28] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1379–1388.

[29] K. Zhan, N. Chen, S. V. N. Santhosh Kumar, G. Kibalya, P. Zhang, and H. Zhang, "Edge computing network resource allocation based on virtual network embedding," *International Journal of Communication Systems*, vol. n/a, no. n/a, p. e5344.

[30] G. Wu, Z. Lin, J. Han, L. Liu, G. Ding, B. Zhang, and J. Shen, "Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval." in *IJCAI*, 2018, pp. 2854–2860.

[31] Z.-D. Chen, C.-X. Li, X. Luo, L. Nie, W. Zhang, and X.-S. Xu, "Scratch: A scalable discrete matrix factorization hashing framework for cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2262–2275, 2019.

[32] Q.-Y. Jiang and W.-J. Li, "Discrete latent factor model for cross-modal hashing," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3490–3501, 2019.

[33] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3027–3035.

[34] 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*, vol. 24, pp. 466–479, 2022.

[35] J. Wu, X. Xie, L. Nie, Z. Lin, and H. Zha, "Reconstruction regularized low-rank subspace learning for cross-modal retrieval," *Pattern Recognition*, vol. 113, p. 107813, 2021.

[36] P.-F. Zhang, J. Duan, Z. Huang, and H. Yin, *Joint-Teaching: Learning to Refine Knowledge for Resource-Constrained Unsupervised Cross-Modal Retrieval*. New York, NY, USA: Association for Computing Machinery, 2021, p. 1517–1525.

[37] R. Xu, C. Li, J. Yan, C. Deng, and X. Liu, "Graph convolutional network hashing for cross-modal retrieval." in *Ijcai*, 2019, pp. 982–988.

[38] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[40] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.

[41] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.

Youxiang Duan professor, received the B.S. degree from the College of Computer and System Science, Nankai University in 1986, and the Ph.D. degree from the School of Geosciences, China University of Petroleum (East China) in 2017. He is currently engaged in teaching and scientific research at the College of Computer Science and Technology, China University of Petroleum (East China). The main research directions include machine learning, network and service computing, etc.

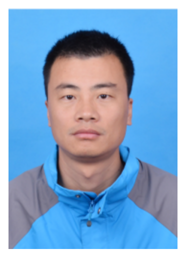
Ning Chen (Graduate Student Member, IEEE) is currently studying for a master's degree in College of Computer Science and Technology, China University of Petroleum (East China).

He has authored several high-level papers, including IEEE WCM, IEEE TVT, PR, KBS, IEEE Access, IJCS, etc. He also serves as a reviewer for several journals/conferences, such as PR, IEEE T-ITS, IEEE Access, CSAE, etc. His research interests include cross-modal retrieval, network virtualization, and future network architecture.

Ali Kashif Bashir (Senior Member, IEEE) is currently a Reader with the Department of Computing and Mathematics, Manchester Metropolitan University, U.K. He is also affiliated with the University of Electronic Science and Technology of China (UESTC), China; the National University of Science and Technology, Islamabad (NUST), Pakistan; and the University of Guelph, Canada, in honorary roles. He has published over 200 research articles, delivered more than 30 invited talks across the globe, organized more than 40 guest editorials, and chaired around 35 conferences and workshops. Since 2016, he has been serving IEEE Technology, Policy and Ethics Newsletter as an Editorin-Cheif. He is also an Associate Editor of several IEEE, Springer, IET, and MDPI journals.

Mohammad Dahman Alshehri (Member, IEEE) received the Ph.D. degree in artificial intelligence of cybersecurity for the Internet of Things (IoT) from the University of Technology Sydney (UTS), Australia. He is currently an Assistant Professor with the Computer Science Department, Taif University, Saudi Arabia, and a Visiting Professor with the School of Computer Science, UTS. Furthermore, he has published several publications in highranked international journals, top-tier conferences, and chapters of book. His main current research interests include cybersecurity, artificial intelligence, the Internet of Things (IoT), and trust and reputation. He has received number of international and national awards and prizes.

Lei Liu (Member, IEEE) received the B.Eng. degrees in communication engineering from Zhengzhou University, Zhengzhou, China, in 2010. He received the M.Sc. and Ph.D degrees in communication engineering from Xidian University, Xi'an, China, in 2013 and 2019, respectively. From 2013 to 2015, he worked in a technology company. From 2018 to 2019, he was supported by China Scholarship Council (CSC) to be visiting Ph.D Student in University of Oslo, Norway. He is currently a lecture with the Department of Electrical Engineering and Computer Science in Xidian University. His research interests include, vehicular ad hoc networks, intelligent transportation, mobile edge computing and Internet of Thing.



Peiyiing Zhang (Member, IEEE) is currently an Associate Professor with College of Computer Science and Technology, China University of Petroleum (East China). He received his Ph.D. in School of Information and Communication Engineering at University of Beijing University of Posts and Telecommunications in 2019. Dr. Zhang has published multiple IEEE/ACM Trans./Journal/Magazine papers, such as IEEE TII, IEEE T-ITS, IEEE TVT, IEEE TNSE, IEEE TNSM, IEEE Network, IEEE IoT-J, IEEE COMMUN MAG, etc. He served as the Technical Program Committee of IEEE ICC'23, IEEE ICC'22, DPPR 2021, ICIST2022, Globecom 2022, Globecom 2021, COMNETSAT 2020, ICICoS 2022, SoftIoT 2021, IWCMC-Satellite 2020, and IWCMC-Satellite 2022, etc. He is the Leading Guest Editor of Electronics, Wireless Communications and Mobile Computing, Mathematical Problems in Engineering, Frontiers in Psychiatry, and the editorial board of Artificial Intelligence and Applications (AIA). His research interests include semantic computing, future internet architecture, network virtualization, and artificial intelligence for networking.

Keping Yu (Member, IEEE) received the M.E. and Ph.D. degrees from the Graduate School of Global Information and Telecommunication Studies, Waseda University, Japan, in 2012 and 2016. He was a Research Associate and a Junior Researcher with the Global Information and Telecommunication Institute, Waseda University, from 2015 to 2019 and from 2019 to 2020, where he is currently a Researcher. His research interests include smart grids, information-centric networking, Internet of Things, blockchain, and information security. He was the Chair of IEEE/CIC ICCC, IEEE VTC2020-Spring, SCML2020, MONAMI 2020, CcS2020, and ITU Kaleidoscope 2016. He has served as a TPC Member for more than ten international conferences. He has been a Lead Guest Editor for Sensors, Peer-to-Peer Networking and Applications, and Energies, and a Guest Editor for IEICE Transactions on Information and Systems, Intelligent Automation and Soft Computing, and Computer Communications. He is an Editor of the IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.