Information Processing and Management 59 (2022) 102743

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Cognitive multi-modal consistent hashing with flexible semantic transformation

Junfeng An [a], Haoyang Luo [a], Zheng Zhang [a,*], Lei Zhu [b], Guangming Lu [a]

[a] Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen 518055, China
[b] School of Information Science and Engineering, Shandong Normal University, Jinan 250000, China

ARTICLE INFO

Keywords:
Social geo-media
Learning to hash
Semantic preserving
Discrete optimization
Similarity search

ABSTRACT

Multi-modal hashing can encode the large-scale social geo-media multimedia data from multiple sources into a common discrete hash space, in which the heterogeneous correlations from multiple modalities could be well explored and preserved into the objective semantic-consistent hash codes. The current researches on multi-modal hashing mainly focus on performing common data reconstruction, but they fail to effectively distill the intrinsic and consensus structures of multi-modal data and fully exploit the inherent semantic knowledge to capture semantic-consistent information across multiple modalities, leading to unsatisfactory retrieval performance. To facilitate this problem and develop an efficient multi-modal geographical retrieval method, in this article, we propose a discriminative multi-modal hashing framework named Cognitive Multi-modal Consistent Hashing (CMCH), which can progressively pursue the structure consensus over heterogeneous multi-modal data and simultaneously explore the informative transformed semantics. Specifically, we construct a parameter-free collaborative multi-modal fusion module to incorporate and excavate the underlying common components from multi-source data. Particularly, our formulation seeks for a joint multi-modal compatibility among multiple modalities under a self-adaptive weighting manner, which can take full advantages of their complementary properties. Moreover, a cognitive self-paced learning policy is further leveraged to conduct progressive feature aggregation, which can coalesce multi-modal data onto the established common latent space in a curriculum learning mode. Furthermore, deep semantic transform learning is elaborated to generate flexible semantics for interactively guiding collaborative hash codes learning. An efficient discrete learning algorithm is devised to address the resulting optimization problem, which obtains stable solutions when dealing with large-scale multi-modal retrieval tasks. Sufficient experiments performed on four large-scale multi-modal datasets demonstrate the encouraging performance of the proposed CMCH method in comparison with the state-of-the-arts over multi-modal information retrieval and computational efficiency. The source codes of this work could be available at https://github.com/JunfengAn1998a/CMCH .

1. Introduction

Recent years have witnessed the explosive growth of social media with the accumulation of an immense amount of users' information and interactions. Within these daily online social interactions, there are amounts of geo-media data, e.g. tweets with

* Corresponding author.
E-mail addresses: JunfengAn1998a@gmail.com (J. An), haoyangluo111@gmail.com (H. Luo), darrenzz219@gmail.com (Z. Zhang), leizhu0608@gmail.com (L. Zhu), luguangm@hit.edu.cn (G. Lu).

https://doi.org/10.1016/j.ipm.2021.102743
Received 28 April 2021; Received in revised form 19 August 2021; Accepted 29 August 2021
Available online 24 September 2021

images of a scenic spot and its description, which leads to large-scale multi-modal datasets. Meanwhile, the retrieval of this geo-related information is a high-frequency requirement of citizens for their tours or communications, and the geographical information need of end-users are not only restricted to a single modality (Kumar, Heuten, & Boll, 2013; Shen et al., 2021) but multi-sources data. Therefore, it is a significant topic for researchers to develop an efficient multi-modal geographic information retrieval (GIR) method to process and manage these vast amounts of Multimedia data.

These multimedia data (such as images, texts, audios and videos) (Wang, Zhang, Luo, Huang, & Shen, 2020; Wang, Zhang, song, Sebe, & Shen, 2018; Zheng et al., 2019; Zhu et al., 2021) dramatically challenges the efficiency and efficacy of the modern information search engines (Cao, Feng, Lin, Cao, & He, 2020). As such, developing lightweight yet effective similarity search techniques is of permanent importance to extract and retrieve valuable information from an ever-growing volume of big data with high dimensionality. Due to the outstanding efficiency of computation and storage, hashing has been recognized as one of the most popular techniques for learning compact binary representations from the high-dimensional features (Zhang et al., 2019; Zheng et al., 2019). Owing to the promising capabilities on balancing the efficiency and performance, hashing can satisfy the GIR end-users' demand for efficiency and has gained considerable attention from the academic and industrial communities, and extensive hashing-based retrieval methods have been proposed to accelerate the retrieval efficiency and promote their search accuracy (Zhang, Lai et al., 2019; Zhang et al., 2018; Zhang, Liu, Shen, Shen, and Shao, 2019).

The early attempts on hashing mainly focus on uni-modal hashing-based retrieval, especially on large-scale image retrieval (Gong, Lazebnik, Gordo, & Perronnin, 2013; Liu, Wang, Ji, Jiang, & Chang, 2012b). Based on whether the samples involved in learning process, they are generally divided into two categories, *i.e.*, data-independent methods (Raginsky & Lazebnik, 2009) and data-dependent ones (Weiss, Torralba, Fergus, et al., 2008). Extensive studies have demonstrated that data-dependent methods (also known as learning to hash) can generate more reliable hash codes than those produced by the data-independent methods, because these hash codes are derived from data rather than a simple random projection. Typically, according to the use of labels, the existing data-dependent hashing methods could be grouped into unsupervised and supervised hashing methods (Liu, Wang, Kumar, & Chang, 2011; Weiss et al., 2008).

Typically, the unsupervised hashing methods mainly explore the underlying distribution structures or topological metrics to preserve them into the learned compact hash codes. For example, the well-known iterative quantization (ITQ) method (Gong et al., 2013) was proposed to minimize the quantization error between the original high-dimensional visual features and the lower-dimensional binary codes by using the simple but effective orthogonal rotation matrix. Although the unsupervised hashing models can provide more feasible and preferable generalization capabilities on new data points, the learned hash codes are incapable of generating discriminative hash codes because the precious labels are ignored in the learning process (Lin, Shen, Suter, & Van Den Hengel, 2013; Liu et al., 2012b; Zhang, Lai et al., 2019). For example, the kernalized supervised hashing (Liu et al., 2011) discovers the relationship between the Hamming distance and inner product, and then develops a pairwise similarity preservation formulation to generate semantic-preserving hash codes. Afterwards, the supervised hashing methods are generally derived by using the discriminative category information in the hash function construction. Due to the promising performances, supervised hashing methods have been intensively studied to promote the discriminant of the learned hash codes, yielding superior results compared to the unsupervised ones.

Generally, when facing the ever-increasing number of social geo-media samples, we always encounter large volume of geographical data collected from different sources, resulting in multi-modal geo-media data. How to effectively derive the uniform binary codes from the multi-modal data becomes one of the most challenges for multi-modal hashing research. In order to deal with massive data analysis, cross-modal retrieval has been emerging to search semantically similar data instances in one modality when given a query sample from the other modality, e.g. given the text information of a category of geographical landscape, returning the semantic similar geographical samples. Technically, the core of cross-modal hashing lies in the common subspace learning between two domains, which could not be used for the situations that are more than two modalities. Moreover, cross-modal hashing focuses on cross domain learning, but how we could jointly employ the multi-modal geographical information for robust common hash codes learning is not considered. Many researches on multi-modal learning have attracted considerable attentions in feature learning and information retrieval (Lu, Zhu, Cheng, Nie, & Zhang, 2019; Zhu, Lu, Cheng, Li, & Zhang, 2020).

Recently, incorporating multi-modal data for more accurate information retrieval has attracted tremendous enthusiasm because multiple knowledge is provided for each instance, and their description is also greatly enriched. Importantly, due to the variations of data structure and feature distribution among multiple modalities, semantic consistent learning (Lu et al., 2019; Zheng et al., 2019; Zhu et al., 2020) becomes the mainstream of multi-modal learning. For example, multiple feature hashing (MFH) (Song, Yang, Huang, Shen, & Luo, 2013) integrates the locality similarity preservation and multiple graph construction into one unified learning framework. Multi-view latent hashing (MVLH) (Shen, Shen, Sun, & Yuan, 2015) builds a multi-view latent subspace derived from the shared kernel feature space for learning latent common hash codes. Many multi-modal hashing methods (Liu, He, Liu, & Lang, 2012a; Liu, Yu, & Shao, 2015; Shen et al., 2018, 2015; Song et al., 2013; Yang, Shi, & Xu, 2017) have been developed to combine the advantages of multi-modal knowledge for effective hash codes learning.

Although some multi-modal hashing models have been proposed to alleviate multi-modal inconsistent gap, these methods suffer from some deficiencies that need to be tackled. When learning common latent representations from multi-modal data, the existing methods tend to learn modality-specific features followed by a simple concatenation or equally project multi-modal data onto the same Hamming space. However, different modalities always contain independent physical characteristics, and these strategies are incapable of capturing the heterogeneous complementary properties of multiple modalities, leading to inferior binary codes. Moreover, the semantic labels are always used for classification or pairwise similarity construction, which is incapable to fully explore the intrinsic semantics embedded in multiple modalities, leading to suboptimal semantic hash codes. Furthermore, the

robustness of hash function learning is less considered in the existing multi-modal space construction, and how to cognitively generate consensus data space in an adaptive learning manner is of great importance in multi-modal hashing. Additionally, the discrete optimization is still under-explored to boost the effectiveness of the learned hash codes.

To overcome the above deficiencies and develop an efficient multi-modal GIR method, in this paper, we propose a discriminative multi-modal hashing framework, dubbed Cognitive Multi-modal Consistent Hashing (CMCH), which can progressively formulate the structure-consistent latent space over heterogeneous multi-modal data and jointly explore the informative semantics based on the deep transform learning strategy. Specifically, an adaptive weighting scheme is formulated to explore the heterogeneous correlations of the samples from multiple modalities, such that the underlying heterogeneous characteristics of data could be adaptively determined and integrated in an self-weighing manner (Liang, Shen, Han, Lei, & Ling, 2017). Moreover, the self-paced learning strategy is leveraged to cognitively aggregate multiple modality features into a well-established latent space. Furthermore, a well-designed deep semantic transformation, for the first time, is employed to recover the informative semantic information and builds the semantic-aware hash codes. Because the commonly-used discrete cyclic coordinate descent (DCC) optimizes code bit-by-bit relying on dissimilar pairs (Wang, Luo, & Xu, 2020), it leads to an unacceptable time-consuming for online application. An efficient discrete optimization algorithm is derived to solve the resulting optimization problem. Finally, considering social multi-modal data cannot be collected one-time due to the limitation of time and space in the application (He, Du, Zhuang, Yin, & Long, 2020), we train the online hash function to convert out-of-sample queries into hash codes. Extensive experiments validate the superiority of the proposed learning framework on different datasets. The main contributions of this work are summarized as follows.

1. We propose a discriminative multi-modal GIR hashing framework, called Cognitive Multi-modal Consistent Hashing (CMCH), to achieve collective multi-modal information fusion for learning high-quality hash codes. Our CMCH jointly considers structure-consistent latent space construction, deep semantic transformation and cognitive feature learning in one unified framework.
2. We develop an efficient parameter-free collaborative multi-modal learning scheme to adaptively excavate the heterogeneous geographical information embedded in the multiple modalities of geo-media, and the common latent space is progressively constructed based on the human cognition, that is, learning from easy to hard.
3. We reconstruct an optimized informative semantic space based on the deep semantic transformation to further enhance the semantic correlations between the semantic space and generated discrete hash space, such that the generated semantic-preserving hash codes belonging to the same geographical conceptual class share similarities but separate themselves from other categories, which can significantly improve accuracy of GIR.
4. We derive a fast discrete learning algorithm to solve the resulting discrete optimization problem to achieve minimum quantization error. Moreover, extensive experimental results on four benchmark datasets demonstrate the efficiency and effectiveness of our CMCH over the state-of-the-art multi-modal hashing methods.

The remaining parts of this work are organized as follows. Section 2 introduces some related works. Section 3 details the objective formula of the proposed CMCH and some theoretical analyses. The extensive experiments are presented in Section 4. Section 5 compares and analysis the experimental results followed by the conclusive remarks in Section 6.

2. Related work

In this section, we systematically review some related topics on learning to hash, *i.e.*, geographic information retrieval methods, uni-modal hashing and multi-modal Hashing.

2.1. Geographic information retrieval

Over the years, geolocation information is getting easier to obtain with the popularization of smart devices, and geographic multimedia retrieval has been widely concerned. Some geographical multimodal information retrieval methods have been proposed. For instance, Ref. Dang-Nguyen, Boato, Moschitti, and Natale (2012) proposed to automatically learn the fusion weight of images with tags and GPS modalities in a supervised approach based on Support Vector Machines (SVMs). But it directly utilizes original data rather than representative fused latent features. Therefore, this strategy results in inferior performance. Moreover, Ref. Purificato and Rinaldi (2018) proposed an integration system to prove the association of geographic data in order to improve the performance of multi-modal retrieval tasks.

2.2. Uni-modal hashing

Uni-modal hashing generally performs binary representation learning and retrieval samples on the same uni-modal data. Notably, the existing uni-modal hashing methods are roughly grouped into two subfields, *i.e.*, shallow uni-modal hashing methods and deep uni-modal hashing methods. The shallow models compress the high-dimensional hand-crafted features into lower-dimensional binary codes, meanwhile preserving their distribution or discriminative semantic properties into the learned hash codes. For example, spectral hashing (Weiss et al., 2008) formulates the subspace consistency between pairwise samples and builds the relationship between the high-dimensional visual features and the compact hash codes. To reduce the computational complexity of graph construction, anchor graph hashing (Liu et al., 2011) gives a feasible solution to construct light-weight graph by selecting anchor samples as the intermediary points. Moreover, some methods embeds the semantic labels into the hash codes learning process. For

example, ITQ (Gong et al., 2013) with canonical correlation analysis (CCA) introduces the supervised CCA into the quantization procedure for supervised hashing. Kernelized supervised hashing (Liu et al., 2012b) is the first pairwise hashing framework to learn discriminative hash codes. The supervised discrete hashing integrates the hash function learning and linear regression into a unified learning model.

Benefiting from the advancement of deep learning, deep hashing becomes popular to jointly learn nonlinear hash functions and hash codes in an end-to-end manner. Generally, deep hashing models can be achieve state-of-the-art hashing performance. For example, Li, Wang, and Kang (2016) proposed a deep supervised hashing with pairwise labels method, which derives the deep pairwise hashing module under a probabilistic Bayesion inference. Shen et al. (2018) constructed an unsupervised deep hashing architecture by using anchor graph learning and alternating discrete optimization. Li, Sun, He, and Tan (2017) incorporated the deep convolutional neural networks, pairwise label preservation and category-level classification into one hash codes learning framework for generating effective hash codes.

However, the aforementioned uni-modal hashing methods fail to handle the complex information retrieval on samples derived from multiple modalities. To enable information retrieval among multiple modalities, one intuitive idea is to directly concatenate features on different modalities for multi-modal feature learning, while such simple merging strategy may increase the information redundancy and fail to full exploit the complementary semantic association between modalities, leading to information loss and inferior performance.

2.3. Multi-modal hashing

The main objective of multi-modal hashing (Kim & Choi, 2013; Liu et al., 2015; Shen, Shen et al., 2018; Song et al., 2013) is to integrate multi-modal data into a unified common space for hash codes learning. Meanwhile, the semantic consistency and heterogeneous information are jointly preserved into the learned optimal hash codes.

Motivated by the unsupervised uni-modal hashing, the pioneering multi-modal hashing, *i.e.,* composite hashing with multiple information sources (CHMIS) (Zhang, Wang, & Si, 2011), integrates multiple graphs on modality-specific features to make locality similarity preservation. Multiple feature hashing (MFH) (Song et al., 2013) was designed to extract multiple features on video and simultaneously employ local and global structures to learn semantic hash codes. Multi-view alignment hashing (MAH) (Liu et al., 2015) constructs multiple non-negative matrix factorization modules with multi-graph regularization to learn shared hash codes in a subspace learning manner. Multi-view discrete hashing (MvDH) (Shen, Shen et al., 2018) takes the pseudo labels produced by k-means as supervision to generate discriminative hash codes. Multi-view anchor graph hashing (MVAGH) (Kim & Choi, 2013) constructs multiple view-specific anchor graphs and integrates them into multi-view anchor graphs for similarity modeling. Multi-view latent hashing (MVLH) (Shen et al., 2015) adopts a latent subspace learning strategy to build latent hash codes based on an adaptive weighting manner, which is determined by the reconstruction error within each view-specific features.

Recently advanced multi-modal hashing methods are generally built on supervised hashing schemes. For example, flexible multi-modal hashing (Zhu et al., 2020) was recently proposed to fuse multiple features into a joint hash space under a pairwise similarity preserving framework. Online multi-modal hashing with dynamic query-adaption (OMH-DQ) (Lu et al., 2019) incorporates latent hash codes learning and pairwise similarity preserving into one unified hash codes learning framework. Fast discrete multi-view hashing (Liu, Zhang, & Huang, 2020) utilizes multi-view semantic dictionary learning and the correntropy-induced optimization for discriminative hash codes learning. Fast discrete collaborative multi-modal hashing (FDCMH) (Zheng et al., 2019) jointly considers asymmetric hashing and common latent space construction in one learning model.

Different from the existing methods, this paper proposes a cognitive multi-modal learning framework, which learns the unified hash codes under an adaptive multi-modal common space construction in conjunction with deep semantic transformation. In this way, the modality correlation and semantic reconstruction are concurrently discovered in a human cognitive learning scheme.

3. Our proposed method

To ensure the logical integrity of the article, in this section, we illustrate the standard notation and problem definition in the first place. Then discuss details of every proposed module following by the theoretical analysis of the whole framework. Specifically, we firstly introduce a similarity-preserved mapping method with linear complexity to generate syncretic data with randomly selected anchors to fill the semantic gaps and excavate the intrinsic relationships between different modalities. Secondly, we construct a latent subspace to learn unified hash codes for bridging the inconsistent semantic gaps across different modalities. In addition, a flexible semantic supervision module is introduced to preserve specific category-level information. Moreover, we develop a robust cognition learning strategy to strengthen the model's ability to eliminate noise and outliers. Eventually, the upper modules are synthetically devised into a unified loss function. Then an introduction of optimization is given to iteratively update each parameter by fixing other parameters as the current state to minimize each term of the loss, including the novel discrete "one-step" hash method. Finally, we extend our method to an online scenario for unseen data.

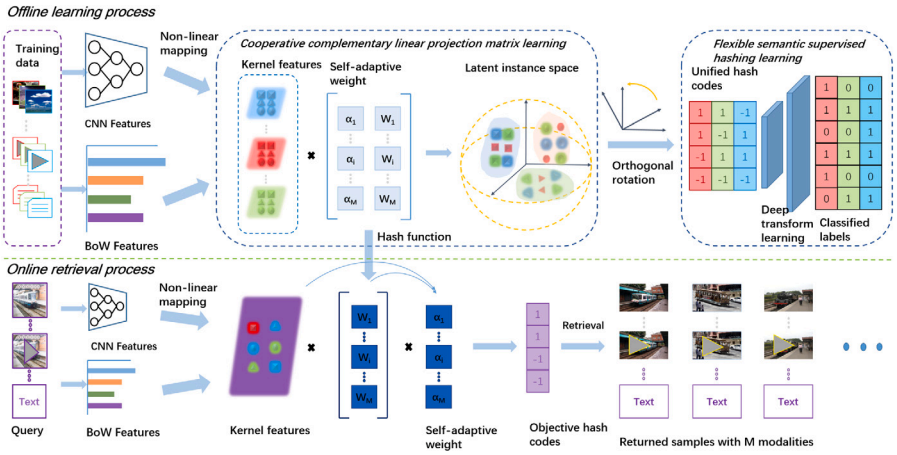


Fig. 1. The overall framework of our proposed CMCH method. In the offline learning process, features of each modality are mapped into a unified structure by random anchors in the RBF space, and the objective hash codes are cognitively learned under the supervision of flexible semantic projection. In the online process, the retrieval hash codes of multi-modal query samples are adaptively generated by the learned hashing function.

3.1. Notation and problem definition

This work focuses on researching the supervised multi-modal hashing retrieval and improving its robustness of semantic similarity search. Without loss of generality, we use bold capital letters to represent matrix, e.g. $\mathbf{A}$. $\mathbf{A}_{i,j}$ denotes the (i,j)-element of $\mathbf{A}$. Moreover, bold lower-case letters represent column vectors, e.g. $\mathbf{a}$. $diag(\mathbf{a})$ returns a square diagonal matrix with vector $\mathbf{a}$ as the main diagonal. $sgn(\mathbf{A}) \in \{-1, 1\}$ is an element-wise sign operation on the matrix $\mathbf{A}$. $\mathbf{A}^{\mathbf{T}}$ is the transposed matrix of $\mathbf{A}$, and $logdet(\mathbf{A})$ is the logarithm of determinant $\mathbf{A}$. $\sqrt{\mathbf{A}}$ is the element-wise square root matrix of $\mathbf{A}$. Frobenius norm of $\mathbf{A}$ is represented as $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}\mathbf{A}^{\mathbf{T}}) = tr(\mathbf{A}^{\mathbf{T}}\mathbf{A})$ and l_{12} norm is defined as $\|\boldsymbol{A}\|_{12} = (\sum_j \sqrt{\sum_i \mathbf{A}_{ij}^2})$.

A set of n multi-modal data instances with M modalities is denoted as $\mathbf{O} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}\}$. $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_n^{(i)}] \in \mathbb{R}^{d_i \times n}$ is the collection of the ith modality features, where d_i is the dimensionality for the ith modality. In this paper, we assume that different modalities of the same sample x_j share a common semantic label. We have c semantic labels of n samples, which can be denoted as $\mathbf{Y} \in \{0, 1\}^{c \times n}$. For instance, x_j belongs to kth class whose kth entry of $\mathbf{Y}_{\cdot j}$ is '1', *i.e.*, $\mathbf{Y}_{kj} = 1$, otherwise $\mathbf{Y}_{kj} = 0$.

The purpose of this paper is to learn common unified hash codes for multi-modal data, which are written as $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n] \in \{-1, 1\}^{h \times n}$, where $\mathbf{b}_i \in \{-1, 1\}^{h \times 1}$ denotes the hash codes for the ith sample. We aim to learn a set of hashing functions to retain inherent characteristics of data during the learning process of the consistent hash codes. Moreover, the objective model could efficiently handle out-of-sample multi-modal queries, for which we can easily obtain a sequence of semantically relevant results of geographical information from large-scale multi-modal databases. The whole learning framework of our CMCH is sketched in Fig. 1.

3.2. Cognitive multi-modal consistent hashing learning

Generally, features of different geo-media modalities have heterogeneous structures. It is necessary for the model to fuse them together in order to fill the semantic gaps and excavate the intrinsic relationships between different modalities. To make features be nonlinearly separable, we introduce an effective similarity-preserve mapping method with linear complexity to generate syncretic data, which serves as the base of binary code learning. Specifically, we randomly select p anchor samples and construct the nonlinear feature embeddings for the ith modality by using the simple Gaussian kernel function. And we map $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_n^{(i)}] \in \mathbb{R}^{d_i \times n}$ into a p-dimensional nonlinear feature vectors, which is defined as $\varphi(X^{(i)}) = [\boldsymbol{\varphi}_1^{(i)}, \boldsymbol{\varphi}_2^{(i)}, \ldots, \boldsymbol{\varphi}_n^{(i)}] \in \mathbb{R}^{p \times n}$. Particularly, $\boldsymbol{\varphi}_k^{(i)} = [exp(-\frac{\|\mathbf{x}_k^{(i)} - \mathbf{a}_1^{(i)}\|^2}{2\sigma^2}), \ldots, exp(-\frac{\|\mathbf{x}_k^{(i)} - \mathbf{a}_p^{(i)}\|^2}{2\sigma^2})]^T$ is the nonlinear feature vector for the kth instance of the ith modality, and $\{\mathbf{a}_j\}_{j=1}^p$ are the randomly sampled p anchor points selected from training samples from different modalities, and σ is the Gaussian kernel width.

3.2.1. Cooperative latent space construction

Practical data are always collected by different sources obtained from different sensors. Based on this, different modalities tend to have independent physical characteristics, *i.e.*, heterogeneous properties. Since such heterogeneous data features are used for illustrating the same subject from different perspectives, exploring shared knowledge embedded in multi-modal data is valuable to construct effective multi-modal hashing function.

In this work, we proposed to construct a latent subspace to learn unified hash codes for bridging the semantic inconsistent gaps across different modalities. To this end, we construct a structure-consistent latent space $\mathbf{V} \in \mathbb{R}^{h \times n}$, where h is the length of objective hash codes. Specifically, each modality is mapped into latent embedding space by a modality-specific linear projection matrix

$\mathbf{W}^{(i)} \in \mathbb{R}^{h\times p}$. Meanwhile, to fully consider the importance of different modalities, we introduce an adaptive modality weighting factor $\alpha^{(i)}$ to identify the significance of each modality in unified feature learning. As such, we have

$$\begin{aligned} &\min_{\alpha^{(i)},\mathbf{W}^{(i)},\mathbf{V}} \| \sum_{i=1}^{M} \alpha^{(i)}\mathbf{W}^{(i)T}\varphi(\mathbf{X}^{(i)}) - \mathbf{V}\|_F^2 + \lambda\|\mathbf{W}^{(i)}\|_{12}, \\ &\quad s.t. \quad \sum_{i=1}^{M} a^{(i)} = 1, 0 < a^{(i)} < 1, \end{aligned} \tag{1}$$

where λ is a balance parameter. Because high dimensional data inevitably includes redundant information, such as repeating data and outliers, we introduce the $\|\mathbf{W}^{(i)}\|_{12}$ regulation term to choose the most representative samples from the training data by imposing structured sparsity constraint. Moreover, we introduce an orthogonal rotation matrix, $\mathbf{R} \in \mathbb{R}^{h\times n}$, as a rotation matrix from the cooperative latent space to the objective binary hash space. Then, the objective function of cooperative latent space construction is formulated as

$$\min_{\mathbf{R},\mathbf{V},\mathbf{B}} \|\mathbf{RV} - \mathbf{B}\|_F^2 \quad s.t. \quad \mathbf{RR}^T = \mathbf{I}, \mathbf{B} \in \{-1, 1\}^{h\times n}. \tag{2}$$

3.2.2. Flexible semantic supervision

Because of its specific classified semantic information, the semantic matrix is the most descriptive composition compared with digital-media modalities. In order to explore the intrinsic semantic embedding and preserve it into the learned hash codes, semantic information is indispensable for the hashing learning model. However, the upper mentioned modules only take advantages of the low-dimensional multi-modal data and the transformation of common latent space to the nearby binary space. They lose the protection of semantic knowledge that includes the most powerful discriminative information.

Although some existing multi-source hashing methods have made use of semantic supervision, they have to pre-construct a huge pairwise matrix including similarities for every pair of samples in the whole large-scale training set, which takes up considerable storage space. Moreover, this pairwise matrix only preserves pairwise information instead of specific category-level information, losing discriminative classified information and leading to inferior hash codes. Therefore, our model focuses on making use of specific category labels rather than pairwise similarity matrix to supervise the learning process.

In order to flexibly recover classified labels $\mathbf{Y}$ in the hash space and further enhance the correlations between semantic space and the objective binary hash space without pre-constructed pairwise supervised matrix. We introduce a well-designed deep transformation strategy that distills original semantic labels into binary codes. By this means, the generated binary codes belonging to the same category are compacted and simultaneously separated well from others belonging to different categories. The deep projection is formulated as

$$\mathbf{Z}_N = \mathbf{P}_N(\mathcal{T}...(\mathbf{P}_2(\mathcal{T}(\mathbf{P}_1\mathbf{Z}_0)))), \tag{3}$$

where $\mathcal{T}$ is the activation function of every layer set as soft-max function in this work, without which all of the $\mathbf{P}_g$ will collapse into a single one. The gth layer is easily formulated as the standard transform learning

$$\begin{aligned} &\min_{\mathbf{P}_g,\mathbf{Z}_g} \mathcal{J}(\mathbf{P}_g, \mathbf{Z}_g) = \|\mathbf{P}_g\mathbf{Z}_{g-1} - \mathbf{Z}_g\|_F^2 + \mu(\|\mathbf{P}_g\|_F^2 - logdet(\mathbf{P}_g)), \\ &\quad s.t. \quad \mathbf{Z}_0 = Y, \mathbf{Z}_N = B, \end{aligned} \tag{4}$$

where the $-logdet(\mathbf{P}_g)$ imposes a full rank on the learned transformation matrix to prevent trivial solutions, and N is the depth of the transform learning network. Every atom of $\mathbf{Z}_g$ can be seen as a local filter that maps the corresponding classified labels to the objective hashing representation $\mathbf{B}$, which further improves the representation ability of the objective hash codes. This flexible learning structure guarantees the subtle valuable information to be captured in the offline hash codes learning process. The $\|\mathbf{P}_g\|_F^2$ serves as a penalty to keep the $-logdet(\mathbf{P}_g)$ from overly increasing and deteriorate the results, and μ is a balance parameter to prevent overfitting.

3.2.3. Robust cognition learning strategy

There are generally unavoidable noise and outliers in large-scale datasets. However, the mainstream of the existing multi-modal hashing learning models neglect to filter these invalid data in the learning stage. The commonly-used strategy to decrease their influence is to run the algorithm many times with randomly selected initialization. In this work, we seek to strengthen the model's ability to eliminate noise and outliers when generates the consensus latent space.

Inspired by curriculum learning strategy in Bengio, Louradour, Collobert, and Weston (2009), which introduces a tendency of selecting training samples from 'easy' to 'complex', self-paced learning (SPL) (Guo, Chang, & Zhu, 2020) has made significant progress and provided a feasible solution to construct a cognitive learning strategy for our problem. In particular, SPL tends to select typical samples that provide the most representative features in each modality to learn in the first place. For example, a pair of instances from the same modality could have semantically consistent target information. Moreover, they share their own private information that weakens the target one. Typically, if given two images of visual modality, models with SPL would firstly choose the one that has target instance in the main part to get useful features, and then use these learned features to extract features from the other one with the same but inconspicuous instance. Because of the enhancement of cognitive competence, this learning sequence leads to more precise and robust results than the commonly-used tricks with random and equal selection. Under this

progressive sample selection training strategy, the conceived algorithm can optimize the representation of the modality projection matrix iteratively and enhance the robustness of the learning system. Based on the popular SPL strategy (Guo et al., 2020), we have the following objective function:

$$\begin{aligned} &\min_{\alpha^{(i)},\mathbf{W}^{(i)T},\mathbf{V}} \|(\sum_{i=1}^{M}\alpha^{(i)}\mathbf{W}^{(i)T}\varphi(\mathbf{X}^{(i)})-\mathbf{V})diag(\sqrt{\boldsymbol{r}^{(i)}})\|_F^2+f(\eta,\mathbf{r}),\\ &\quad s.t.\quad \sum_{i=1}^{M}a^{(i)}=1, 0<a^{(i)}<1, \end{aligned} \tag{5}$$

where $diag(\sqrt{\boldsymbol{r}^{(i)}})$ is the cognitive vector to automatically select the learning sequence, and $f(\eta,\mathbf{r})$ is the self-paced regularization. η is a scalar that controls the learning rate. By increasing the penalty iteratively, during the process of optimization, more and more samples are progressively chosen from the training set. With η and other variables fixed in each iteration, $\mathbf{r}^{(i)}$ is calculated based on the adaptive difficulty-level judgment in each iteration. Following the adaptive learning strategy in Xu, Tao, and Xu (2015), we have

$$f(\eta,\mathbf{r})=\sum_{i=1}^{M}(1+e^{-\eta}-\mathbf{r}^{(i)})ln(1+e^{-\eta}-\mathbf{r}^{(i)})+\mathbf{r}^{(i)}lnr^{(i)}-\eta\mathbf{r}^{(i)}. \tag{6}$$

3.2.4. Overall objective function

Integrating every module into a unified framework, we get the objective function of our CMCH as follows:

$$\begin{aligned} \min_{\alpha^{(i)},\mathbf{r}^{(i)},\mathbf{W}^{(i)},\mathbf{V},\mathbf{R},\mathbf{B},\mathbf{P}_g} &\|(\sum_{i=1}^{M}\alpha^{(i)}\mathbf{W}^{(i)}\varphi(\mathbf{X}^{(i)})-\mathbf{V})diag(\sqrt{\mathbf{r}})\|_F^2+\beta\|\mathbf{RV}-\mathbf{B}\|_F^2\\ &+\theta\|\mathbf{P}_N(\mathcal{T}...(\mathbf{P}_2(\mathcal{T}(\mathbf{P}_1\mathbf{Z}_0))))-\mathbf{Z}_N\|_F^2\\ &+\lambda(\sum_{g=1}^{N}(\|\mathbf{P}_g\|_F^2-logdet(\mathbf{P}_g))+\sum_{i=1}^{M}\|\mathbf{W}^{(i)}\|_{12})+f(\eta,\mathbf{r}),\\ s.t.\quad &\sum_{i=1}^{M}a^{(i)}=1, 0<a^{(i)}<1, \mathbf{B}\in\{-1,1\}^{h\times n}, \mathbf{RR}^T=\mathbf{I},\\ &\mathbf{Z}_0=Y, \mathbf{Z}_N=B, \end{aligned} \tag{7}$$

where β, θ, λ are balance parameters. The first term is to progressively explore the complementary information of diverse modalities and construct cooperative latent space with the cognitive learning scheme. Then the second term is to convert the consensus latent space to the objective binary hash space. The third and fourth terms are designed to map the classified information from semantic labels into the expected hash space as a flexible supervision. The last term is the regularization of self-paced learning.

3.3. Optimization algorithm

In this subsection, we introduce an efficient alternative optimization algorithm *i.e.*, optimize each variable in a step when fixing others to solve the resulting problems that iteratively minimizes the objective loss function.

It is notable that problem Eq. (7) includes a discrete constraint on $\mathbf{B}\in\{-1,1\}^{h\times n}$, leading to a NP-hard problem. Most of the existing methods (Kang, Kim, & Choi, 2012; Kim & Choi, 2013; Liu et al., 2015; Song et al., 2013; Zhang et al., 2011) utilize "relaxing+ rounding" strategy to optimize it, that is, relaxed continuous solutions followed by a simple thresholding, which leads to an unavoidable large quantization loss. In addition, another way is to optimize it by a bit-by-bit optimization manner (Shen, Shen et al., 2018), but its time consumption is unacceptable for large-scale data applications. The computational details of each step are described as follows.

$\mathbf{W}^{(i)}$-Step: Fixing other variables, we can optimize modality-specific linear projection $\mathbf{W}^{(i)}$ sequentially. Following the optimization in Nie, Huang, Cai, and Ding (2010), we introduce an auxiliary diagonal matrix $\mathbf{Q}$, where $\mathbf{Q}_{kk}=\frac{1}{\sqrt{\|\mathbf{W}_{.k}\|_2^2+\epsilon}}$ and ϵ is an auxiliary variable to solve the l_{12}-norm regularized problem. The optimization problem *w.r.t.* $\mathbf{W}^{(i)}$ in Eq. (7) is reduced to the following problem:

$$\begin{aligned} &\min_{\alpha^{(i)},\mathbf{r},\mathbf{W}^{(i)},\mathbf{V}} \psi(\mathbf{W}^{(i)})=\|(\alpha^{(i)}\mathbf{W}^{(i)}\varphi(\mathbf{X}^{(i)})+\mathbf{C})diag(\sqrt{\mathbf{r}})\|_F^2+\lambda tr(\mathbf{W}^{(i)}\mathbf{Q}\mathbf{W}^{(i)T}),\\ &\quad s.t.\quad \sum_{i=1}^{M}a^{(i)}=1, 0<a^{(i)}<1, \end{aligned} \tag{8}$$

where $\mathbf{C}=\sum_{j=1,j\neq i}^{M}\alpha^{(j)}\mathbf{W}^{(j)}\varphi(\mathbf{X}^{(j)})-\mathbf{V}$. This problem can be optimized by setting $\frac{\partial\psi}{\partial\mathbf{W}^{(i)}}=0$, then we have

$$\begin{aligned} &\alpha^2\mathbf{W}^{(i)}\varphi(\mathbf{X}^{(i)})\mathbf{r}\varphi(\mathbf{X}^{(i)})^T+\alpha(\sum_{j=1,j\neq i}^{M}\alpha^{(j)}\mathbf{W}^{(j)}\varphi(\mathbf{X}^{(j)})-\mathbf{V})\varphi(\mathbf{X}^{(i)})^T\\ &+\frac{1}{2}\lambda\mathbf{Q}\mathbf{W}^{(i)}=0. \end{aligned} \tag{9}$$

Therefore, the optimization of $\mathbf{W}^{(i)}$ is given as

$$\mathbf{W}^{(i)} = \frac{-\alpha^{(i)}(\sum_{j=1,j\neq i}^{M}\alpha^{(j)}\mathbf{W}^{(j)}\varphi(\mathbf{X}^{(j)}) - \mathbf{V})\varphi(\mathbf{X}^{(i)})^T}{\alpha^{(i)2}\varphi(\mathbf{X}^{(i)})\mathbf{r}\varphi(\mathbf{X}^{(i)})^T + \lambda\mathbf{Q}}. \tag{10}$$

R- Step: By fixing other variables, the optimization problem *w.r.t.* $\mathbf{R}$ becomes

$$\min_{\mathbf{R}} \|\mathbf{RV} - \mathbf{B}\|_F^2, \quad s.t. \quad \mathbf{RR}^T = \mathbf{I}. \tag{11}$$

Inspired by the definition of the optimal $\mathbf{R}$ in Zhu, Shen, Xie, and Cheng (2017), we can optimize $\mathbf{R}$ by using

$$\mathbf{R} = \mathbf{F}_1\mathbf{F}_2^T, \tag{12}$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are the left and right singular matrices of $\mathbf{VB}^T$.

P- Step: We use the alternating direction method of multipliers (ADMM) to solve this problem. By fixing other variables, we convert the optimization of $\mathbf{P}$ to

$$\begin{aligned}&\min_{\mathbf{P}_g,\mathbf{B}} \|\mathbf{P}_N(\mathcal{T}...(\mathbf{P}_2(\mathcal{T}(\mathbf{P}_1\mathbf{Z}_0)))) - \mathbf{Z}_N\|_F^2 + \mu\sum_{g=1}^{N}(\|\mathbf{P}_g\|_F^2 - logdet(\mathbf{P}_g)),\\ &\quad s.t. \quad Z_0 = \mathbf{Y}, Z_N = \mathbf{B},\end{aligned} \tag{13}$$

where $\mu = \frac{\lambda}{\theta}$. Based on the greedy learning paradigm (Bengio, Lamblin, Popovici, Larochelle, et al., 2007), we can optimize Eq. (13) one layer at a time. Because $\mathbf{B}$ has the discrete constraint, we use substitutions $\mathcal{T}(\mathbf{P}_{g-1}...(\mathbf{P}_2(\mathcal{T}(\mathbf{P_1Z_0})))) = \mathbf{Z}_{g-1}$ just until the final substitution $\mathcal{T}(\mathbf{P}_1\mathbf{Y}) = \mathbf{Z}_1$, then apply our proposed one-step discrete optimization method to optimize $\mathbf{B}$. Specifically, the optimal solution of one layer, similar to Eq. (4), is given by

$$\mathbf{Z}_{g-1}\mathbf{Z}_{g-1}^T + \mu\mathbf{I} = \mathbf{GG}^T, \tag{14a}$$

$$\mathbf{G}^{-1}\mathbf{Z}_{g-1}\mathbf{B}^T = \mathbf{U}\Sigma\mathbf{K}^T, \tag{14b}$$

$$\mathbf{P}_g = 0.5\mathbf{K}(\Sigma + (\Sigma^2 + 2\mu\mathbf{I})^{\frac{1}{2}})\mathbf{U}^T\mathbf{G}^{-1}, \tag{14c}$$

$$\mathbf{Z}_g = \mathcal{T}(\mathbf{P}_g\mathbf{Z}_{g-1}), \tag{14d}$$

where Eq. (14a) is to compute the Cholesky decomposition, the Eq. (14b) is to compute the full SVD operator, and inspired by Ravishankar, Wen, and Bresler (2015) we can optimize $\mathbf{P}$ with the Eq. (14c). Here we use hard thresholding as activation function $\mathcal{T}$.

V- Step: By fixing other variables, the optimization of $\mathbf{V}$ degenerates to the following problem:

$$\min_{\mathbf{V}} \|(\sum_{i=1}^{M}\alpha^{(i)}\mathbf{W}^{(i)}\varphi(\mathbf{X}^{(i)}) - \mathbf{V})diag(\sqrt{\mathbf{r}})\|_F^2 + \beta\|\mathbf{RV} - \mathbf{B}\|_F^2. \tag{15}$$

We can optimize $\mathbf{V}$ by calculating the solution of partial derivative w.r.t. $\mathbf{V}$, and we have

$$\mathbf{V} = \frac{\sum_{i=1}^{M}\alpha^{(i)}\mathbf{W}^{(i)}\varphi(\mathbf{X}^{(i)})diag(\mathbf{r}) + \beta\mathbf{R}^T\mathbf{B}}{(diag(\mathbf{r}) + \beta\mathbf{R}^T\mathbf{R})}. \tag{16}$$

B- Step: We fix other variables to update $\mathbf{B}$, then the problem Eq. (7) w.r.t. $\mathbf{B}$ is converted to the following problem:

$$\begin{aligned}&\min_{\mathbf{B}} \beta\|\mathbf{RV} - \mathbf{B}\|_F^2 + \theta\|\mathbf{P}_N\mathbf{Y} - \mathbf{B}\|_F^2,\\ &\quad s.t. \quad \mathbf{B} \in \{-1, 1\}^{h\times n}.\end{aligned} \tag{17}$$

Notice that $\mathbf{B}$ is the matrix of discrete binary elements, then the problem is equivalent to

$$\begin{aligned}&\min_{\mathbf{B}} tr(\mathbf{B}^T(\beta\mathbf{RV} + \theta\mathbf{P}_N\mathbf{Y})),\\ &\quad s.t. \quad \mathbf{B} \in \{-1, 1\}^{h\times n}.\end{aligned} \tag{18}$$

As such, we can get the closed-form solution of $\mathbf{B}$, *i.e.*,

$$\mathbf{B} = sgn(\beta\mathbf{RV} + \theta\mathbf{P}_N\mathbf{Y}). \tag{19}$$

***α*-Step**: To adaptively update the fusion weights $\boldsymbol{\alpha}$, we fix the other variables. Then the optimized $\boldsymbol{\alpha}$ is given by the following step:

$$\begin{aligned}&\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}\mathbf{H}\boldsymbol{\alpha} - \mathbf{f}^T\boldsymbol{\alpha},\\ &\quad s.t. \quad \sum_{i=1}^{M}a^{(i)} = 1, 0 < a^{(i)} < 1,\\ &\mathbf{H}_{i,j} = tr(\mathbf{W}_i\varphi(\mathbf{X}_i)diag(\mathbf{r})\varphi(\mathbf{X}_i)^T\mathbf{W}_j^T),\\ &\mathbf{f}_i = tr(\mathbf{W}_i\varphi(\mathbf{X}_i)diag(\mathbf{r})\mathbf{V}^T).\end{aligned} \tag{20}$$

Algorithm 1 Optimization for CMCH Framework

Input: Original multi-source data $\mathbf{X}$, Classified matrix $\mathbf{Y}$, hyper-parameters β, θ, λ, number of random anchors p.

Output: Collaboratively complementary linear projection matrix $\mathbf{W}$, cooperative latent space matrix $\mathbf{V}$, uniform target hash code $\mathbf{B}$, modal fusion weighting factor $\boldsymbol{\alpha}$.

Random initialization of $\mathbf{V}$,$\mathbf{W}$,$\mathbf{P}_N$,$\mathbf{B}$,$\mathbf{R}$;
$\alpha^{(i)} = \frac{1}{M}$, $\mathbf{r} = \{1\}^{n\times 1}$;
Calculate $\varphi(\mathbf{X})$;
For t = 1 : T do
  Update $\mathbf{W}^{(i)}$ by solving Eq. (10);
  Update $\mathbf{R}$ by solving Eq. (12);
  Update $\mathbf{P}_g$ by solving Eq. (13);
  Update $\mathbf{V}$ by solving Eq. (16);
  Update $\mathbf{B}$ by solving Eq. (19);
  Update $\boldsymbol{\alpha}$ by solving Eq. (20);
  Update $\mathbf{r}$ by solving Eq. (22);
End.

The above problem is a famous quadratic programming problem formulated as $\min_{\alpha^{(i)}} \frac{1}{2}\alpha^{(i)T}\mathbf{H}\alpha^{(i)} - \mathbf{f}^T\alpha^{(i)}$ in Eq. (20), which can be optimized by the Lagrangian multiplier method.

r- Step: The loss term on samples is given as

$$\mathbf{L} = \sum_{i=1}^{M} \alpha^{(i)}\mathbf{W}^{(i)}\varphi(\mathbf{X}^{(i)}) - \mathbf{V}. \tag{21}$$

When other variables are fixed in the iteration, r_s, s is the serial number of training samples, is optimized by

$$r_s = \frac{1 + e^{-\eta}}{1 + e^{l_s - \eta}}, \tag{22}$$

where η is an auxiliary parameter, and l_s is the loss term of every sample, which is defined as

$$l_s = \|\mathbf{L}_{.s}\|_F^2. \tag{23}$$

3.4. Out-of-sample extension

After obtaining binary training set codes from discrete optimization, we propose a hash function in the online retrieval module that directly maps the unseen query *i.e.*, $\{q^{(m)}\}_{m=1}^{M}$ into compact hash codes. Generally, after learning the unified binary $\mathbf{B}$ by aforementioned optimization steps, feature dimensionality reduction matrix $\mathbf{U}_\upsilon$ for the cooperative latent space $\mathbf{V}$ is given by the following linear regression steps:

$$\|\mathbf{U}_\upsilon^T\mathbf{V} - \mathbf{B}\| + \|\mathbf{U}_\upsilon\|_F^2, \tag{24}$$

where $\mathbf{U}_\upsilon$ is the objective hash function that maps the continuous value data into objective binary hash codes. The closed-form solution of $\mathbf{U}_\upsilon$ is

$$\mathbf{U}_\upsilon = (\mathbf{V}\mathbf{V}^T + \mathbf{I})^{-1}\mathbf{V}\mathbf{B}^T, \tag{25}$$

Most of the existing methods apply fixed modality fusion weights learned by the off-line training that neglects various underlying complementary information of the dynamic queries. To eliminate this limitation, we introduce a self-adaptive weighting strategy for the online querying process. For each query instance $\{q^{(m)}\}_{m=1}^{M}$, its representation in the nonlinear kernel space calculated by the anchor-feature embedding strategy in Section 3.2 written as $\{\varphi(\mathbf{q}^{(m)})\}_{m=1}^{M}$. The self-adaptive weight of the query $\alpha_q^{(m)}$ is given by quadratic programming strategy in Eq. (20). We can obtain its representation in latent space by $\upsilon = \sum_{m=1}^{M}\alpha_q^{(m)}\mathbf{W}^{(i)}\varphi(\mathbf{q}^{(m)})$. Then the corresponding binary codes for the query instance are formulated as

$$\mathbf{b} = sgn(\mathbf{U}_\upsilon^T\upsilon). \tag{26}$$

3.5. Complexity analysis

In this subsection, we analyze the computational and memorial complexity of CMCH. The time complexity for $\mathbf{W}^{(i)}$ calculation is $\mathcal{O}(phn)$. The time consumption to compute $\mathbf{R}$ is $\mathcal{O}(h^2n)$. The time complexity to calculate $\mathbf{V}$ is $\mathcal{O}(hn)$. It takes $\mathcal{O}(phn)$ time to generate $\mathbf{B}$. The time consumption to compute every $\mathbf{P}_i$ is of upper bound $max(h^3)$. Finally, $\mathbf{r}$ produces a time complexity of $\mathcal{O}(phn)$. On the large-scale datasets, $n \gg p, n \gg h$. Therefore, the total time complexity is $\mathcal{O}(n)$. What is more, because our CMCH method is the first to directly utilize the label matrix instead of the pairwise similarity matrix and avoids the graph structure, the storage complexity is also $\mathcal{O}(n)$. This validates that the proposed CMCH method is suitable for large-scale multi-modal data retrieval.

Table 1
General statistics of Wiki, MIRFlicker, NUS-WIDE, and MS COCO.

Datasets	Training size	Retrieval size	Query size	Categories
Wiki (Pereira et al., 2013)	2173	2173	693	10
MIR Flickr (Huiskes & Lew, 2008)	17,772	17,772	2243	24
NUS-WIDE (Chua et al., 2009)	193,749	193,749	2085	21
MS COCO (Lin et al., 2014)	115,525	115,525	7762	80

3.6. Convergence analysis

The efficiency of the algorithm depends largely on the rate of convergence. The theoretical guarantee of the proposed optimization is given in the following theorem.

Theorem 1. *The alternating optimization steps monotonically reduce the objective function at each step until it converges to a local optimal point.*

Because of the summation of various positive norms in the problem Eq. (7), the objective function of CMCH is clearly lower bounded. Moreover, every step of our optimization has a closed-form solution for the corresponding subproblem in each iteration. In this way, if we denote the value of the objective function at the $(t+1)$th iteration by $\mathcal{L}^{t+1}$, we can easily deduce that $\mathcal{L}^{t+1}(\mathbf{W}^{(i)}, \mathbf{R}, \mathbf{P}_g, \mathbf{V}, \mathbf{B}, \boldsymbol{\alpha}, \mathbf{r}) > \mathcal{L}^{t}(\mathbf{W}^{(i)}, \mathbf{R}, \mathbf{P}_g, \mathbf{V}, \mathbf{B}, \boldsymbol{\alpha}, \mathbf{r})$, which is always satisfied. Due to the bounded monotone protocol in our learning algorithm, the proposed optimization algorithm can efficiently converge to a local optimal point within limited iterations. Moreover, we will validate the convergent efficiency of our optimization scheme in experiments.

4. Experiments

In this section, we conduct a group of experimental assessments on four publicly available large-scale multimodal benchmark datasets related to geography and evaluate the performance of our proposed CMCH by comparing different evaluation protocols with baselines, including state-of-the-art methods.

4.1. Large-scale datasets

For fairness, we select datasets and respectively construct training and querying collections as the existing state-of-the-art method OMH-DQ (Lu et al., 2019). As reported in OMH-DQ (Lu et al., 2019), our benchmark datasets also include MIR Flickr (Huiskes & Lew, 2008), NUS-WIDE (Chua et al., 2009) and MS COCO (Lin et al., 2014). Furthermore, we add Wiki (Pereira et al., 2013) relying on Wikipedia to further validate the efficacy of the proposed CMCH method. All of these datasets are geographically relevant and include massive multi-modal geographical data such as scenic spot with its corresponding description as the sample shown in Fig. 2. And the corresponding general information is shown in Table 1. There are plenty of geo-media relevant multimodal samples in each dataset. The details of the selected datasets are given as follows:

Wiki (Pereira et al., 2013) contains 2866 paired text-image samples from Wikipedia, each of which has one label from 10 semantic categories, and only those with the same label are considered to be relevant. The visual contents are converted to 128-dimension scale-invariant feature transformation (SIFT) (Lowe, 1999) histograms, and the text contents are represented by 10-dimension Latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) features. We randomly choose 693 paired samples to form the querying set and remain samples are regarded as the retrieval database.

MIRFlickr-25K (Huiskes & Lew, 2008) comprises 25,000 pairs of multi-view samples gathered from one million images collected via Flickr, each of which is annotated with multi-tags of 24 categories' and the data sharing at least one same label are considered to be neighbors. Following the same procedures as OMH-DQ (Lu et al., 2019), we convert images into 4096-dimensional vectors by the Caffe program of VGG Net and texts features into 1386-dimensional vectors by BoW. We randomly choose 2243 instances from MIRFlickr to construct the querying set and 17,772 ones from the remainder to comprise the training data.

NUS-WIDE (Chua et al., 2009) consists of 269,648 multimedia data collected from a popular social network website associated with multiple concepts. In our experiments, visual and textual features are respectively expressed as 4096-dimensional vectors by VGG Net and 1000-dimensional vectors by BoW. Based on this, we randomly choose 2085 images and their corresponding texts as the querying set and the remaining are treated as training samples.

MS COCO (Lin et al., 2014) is a widely used dataset containing over 300,000 multiple labeled data points, distributed in 80 categories. Only samples sharing at least one concept are considered to be relevant. We utilize VGG to extract image features into 4096-dimensional vectors and convert text features by BoW into 2000-dimensional vectors. Based on this, as reported in OMH-DQ (Lu et al., 2019), we randomly select 7762 instances to form the querying set and 115,525 ones from the remainder to form the training set.

Fig. 2. A geographical image of "Birmingham campaign" with textual "Martin Luther King's presence in Birmingham was not welcomed by all in the black community". sampling from Wiki.

4.2. Evaluation metrics

Three commonly-used standard information retrieval evaluation indicators are introduced in this subsection to evaluate the efficiency of CMCH. For Hamming ranking task, we apply the mean Average Precision (mAP) and topK–precision curve to measure the accuracy. Average precision (AP) for a certain query sample q is a comprehensive evaluation criterion calculated by

$$AP(q) = \frac{1}{NR}\sum_{t=1}^{w} P_q(t)\delta_q(t), \tag{27}$$

where w is the volume of retrieval database, and if the t-th retrieved instance is neighbor of the query q, $\delta_q(t) = 1$ otherwise 0. NR represents the number of relevant items in the retrieval database, and $P_q(t)$ is the precision of the top t retrieval instances.

$$mAP(q) = \frac{1}{Q}\sum_{i=1}^{Q} AP(q_i). \tag{28}$$

What is more, topK–precision curves demonstrate the retrieval performance by reflecting the retrieval precision w.r.t. top-ranked K instances. For both of these two evaluation metrics, a higher score portends the preferable retrieval ability. For the multi-modal retrieval, we utilize the precision–recall curve to demonstrate the balance between precision and recall properties of our proposed CMCH.

4.3. Comparison methods

We experimentally testify our CMCH and 7 baseline methods including 4 unsupervised hashing methods, *i.e.*, Multiple Feature Hashing (MFH) (Song et al., 2013), Multi-view Alignment Hashing (MAH) (Liu et al., 2015), Multi-view Latent Hashing (MVLH) (Shen et al., 2015), Multi-view Discrete Hashing (MvDH) (Shen, Shen et al., 2018) and 3 state-of-the-art supervised hashing methods, *i.e.*, Multiple Feature Kernel Hashing (MFKH) (Liu et al., 2012a), Discrete Multi-view Hashing (DMVH) (Yang et al., 2017), Online Multi-modal Hashing with Dynamic Query-adaption (OMH-DQ) (Lu et al., 2019). The descriptions of algorithms are listed as follows:

1. MFH (Song et al., 2013): This is a well-known unsupervised hashing method which preserves the local structure of each modality and integrates global structure during the optimization.
2. MAH (Liu et al., 2015): The author develops a multi-graph regularized non-negative matrix factorization to explore the joint data distribution and learn the latent semantics for the hash codes learning.
3. MVLH (Shen et al., 2015): Modal-specific features are used in this method to reconstruct instances in the hidden subspace with an adaptive weighting strategy.

Table 2
The mAP of compared methods and our proposed CMCH with 16, 32, 64, and 128 bits hash codes on four datasets.

Methods	Wiki				MIRFlickr-25k				NUS-WIDE				MS COCO			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
MFH (Song et al., 2013)	0.4364	0.4899	0.5216	0.5318	0.5842	0.5844	0.5846	0.5850	0.3272	0.3279	0.3284	0.3287	0.3959	0.3966	0.3972	0.3991
MAH (Liu et al., 2015)	0.1161	0.4127	0.5197	0.5417	0.5818	0.5818	0.5818	0.5875	0.3225	0.3225	0.3225	0.3320	0.3962	0.3962	0.3962	0.3961
MVLH (Shen et al., 2015)	0.4210	0.4628	0.4689	0.5082	0.6571	0.6412	0.6012	0.5988	0.4117	0.3617	0.3406	0.3492	0.4138	0.4230	0.4089	0.4076
(Shen, Shen et al., 2018)	0.4196	0.5468	0.5523	0.5509	0.6696	0.6957	0.7295	0.7291	0.4795	0.4944	0.5047	0.5111	0.3967	0.3973	0.3978	0.3983
MFKH (Liu et al., 2012a)	0.3406	0.2778	0.2491	0.2306	0.6256	0.6298	0.6259	0.6150	0.3661	0.3791	0.3764	0.3740	0.4199	0.3964	0.3960	0.3957
DMVH (Yang et al., 2017)	0.5241	0.5706	0.5891	0.6027	0.5818	0.5818	0.5818	0.5818	0.3225	0.3223	0.3225	0.3225	0.3975	0.3982	0.3974	0.4004
OMH-DQ (Lu et al., 2019)	0.5909	0.5932	0.6189	0.6279	0.6625	0.7636	0.7934	0.8006	0.5886	0.5279	0.6112	0.6561	0.4440	0.4438	0.5358	0.5759
Our method	**0.7282**	**0.7644**	**0.7569**	**0.7657**	**0.8344**	**0.8353**	**0.8587**	**0.8827**	**0.7375**	**0.7668**	**0.7843**	**0.7930**	**0.5150**	**0.6068**	**0.6441**	**0.6737**

4. MvDH (Shen, Shen et al., 2018): MvDH applies pseudo labels generated by spectral clustering to learn objective discrete hash codes.
5. MFKH (Liu et al., 2012a): This method exploits multi-kernel linear combinations to preserve the similarities of multi-sources and generate representative hash codes.
6. DMVH (Yang et al., 2017): DMVH constructs $n \times n$ matrix as a semantic graph to preserve the similarities and corresponding information of pairwise samples.
7. OMH-DQ (Lu et al., 2019): This method introduces a novel equivalent self-adaptive fusing weighting strategy to replace the hand-craft modality weights for each testing query to generate binary codes for the new queries.

4.4. Implementation details

The proposed CMCH has 3 parameters: β, θ and λ as shown in Eq. (7). β is used to coordinate the rotational projection from the cooperative latent space to binary hashing space for reducing information loss between continuous and discrete space. Trade-off parameter balancing the flexible transform learning that extracts and reconstructs semantic supervision into hash codes is written as θ, and λ is a penalty parameter to avoid over fitting of the linear projection and flexible semantic supervision. After analyzing the parameter sensitivity on MIRFlicker, we find that the accuracy is insensitive to the hype-parameter β and the optimal β on MIRFlicker is 0.5. Therefore, for simplicity, we set β as 0.5 on all of the datasets. Moreover, the number of random anchors p is related to the scale of datasets. The best choice of p is equally set to 1000 on MIRFlickr, MS COCO, and NUS-WIDE, and 500 on Wiki. Then we use grid search to obtain the optimal value of every hyper-parameter with the other hyper-parameter fixed as the current estimates until the best performance of CMCH is achieved on Wiki, MIRFlickr, MS COCO, and NUS-WIDE, respectively. The optimal parameters are $\{\beta = 0.5, \lambda = 10^{-3}, \theta = 1, p = 500\}$ on Wiki, $\{\beta = 0.5, \lambda = 10^{-7}, \theta = 10, p = 1000\}$ on MIRFlicker, $\{\beta = 0.5, \lambda = 10^{-1}, \theta = 1, p = 1000\}$ on NUS-WIDE, $\{\beta = 0.5, \lambda = 10^{-1}, \theta = 10, p = 1000\}$ on MS COCO, with fixing the transform learning network depth at two. In order to assure the best balance between efficiency and accuracy as analysis in Section 5.7, the maximum iteration number is set to 15. All of our experiments are conducted on MATLAB R2016b on a standard Windows PC with 2.90 GHz Intel(R) Core(TM) i7-10700 CPU and 32G RAM.

5. Results comparison and efficiency discussion

This section conducts various comparison experiments regarding the upper three frequently-used performance metrics and gives thorough discussions from different aspects. Notably, all algorithms are testified five times with different random initializations, and the average results are recorded. Firstly, we compare the retrieval accuracy based on the evaluation metrics given in Section 4.2. Then, training efficiency and stability are revealed in Section 5.2. Moreover, we design a group of ablation studies of every proposed module to validate their effects respectively. In addition, we demonstrate the outstanding representative ability of the learned hash codes by visualizing the learned features with the t-SNE tool. Finally, Sections 5.6 and 5.7 respectively discuss the parameter sensitivity and experimentally illustrate the algorithm convergence.

5.1. Retrieval accuracy comparisons and discussion

To validate the superiority of CMCH, we compare our method with baselines in respect of retrieval precision and other evaluation methods given in Section 4.2. We directly apply default parameters by following the original literatures or fine-tuning on each

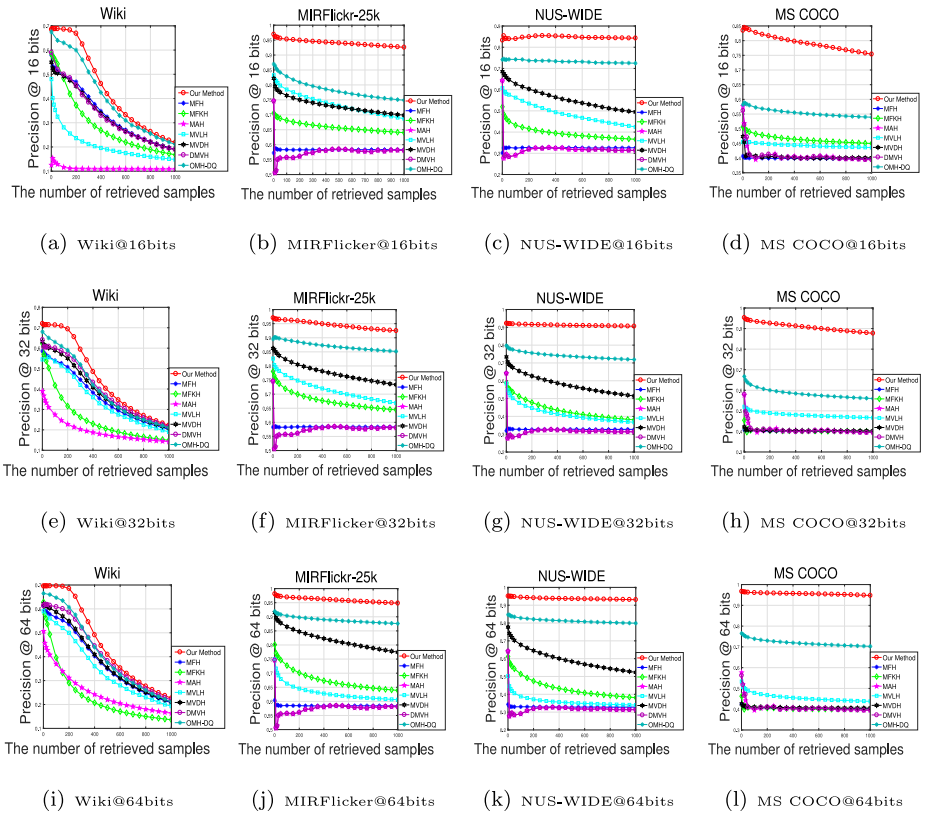


Fig. 3. The precision@topK curves on Wiki, MIRFlicker, NUS-WIDE, and MS COCO with 16, 32, and 64 bits hash codes.

dataset. In order to demonstrate the outstanding performance of CMCH, we successively compare mAP and topK–precision on Hamming ranking task and further compare precision–recall on multi-modal retrieval task.

Table 2 shows the mAP scores comparison of the baseline methods and our CMCH methods on the four datasets with hash code length varies from 16 to 128 bits and the best results are marked with boldface. From the results, we have observations as follows: (1) The proposed CMCH consistently outperforms all of the state-of-the-art supervised hashing and unsupervised hashing baselines, proving its effectiveness in dealing with multimodal retrieval tasks. Notably, compared with OMH-DQ, the most advanced supervised method, the average mAP score of our CMCH on Wiki, MIRFlicker, NUS-WIDE, and COCO2017 is increased by 0.1461, 0.09775, 0.17445, 0.1100, respectively. Moreover, on Wiki, MIRFlicker, NUS-WIDE, and COCO2017, the average mAP of CMCH is 0.2364, 0.1468, 0.2730, and 0.2124 higher than the best unsupervised method MvDH. In addition, even with shorter hash codes, CMCH performs clearly superior to compared methods with longer hash codes. For example, our MIRFlickr-25k mAP score on 16 bits has outperformed the result of the second-best method, OMH-DQ, on 128 bits. (2) We also find that CMCH has progressive performance improvement with the code length increasing, indicating that CMCH can evenly preserve more semantic information in every bit. (3) With the code length increasing, some baseline methods, e.g. MFKH and MVLH, have unstable retrieval ability. The possible reason is their weakness in dealing with the robustness of the learning strategy so that they bring the noise into learned codes leading to unbalanced representative property of each binary bit.

With the increasing retrieval samples, the topK–precision curves of 16, 32, 64 bits hash codes are plotted in Fig. 3. We can easily find that: (1) It is also clear that our method significantly outperforms the other competing methods as the number of samples retrieved increases. (2) This result indicates that our CMCH method can return more relevant samples in the limited results against existing approaches. The superiority of CMCH credits the complete specific semantic-preserve flexible label transform learning in the unified learning framework.

The precision–recall accuracies for multi-modal retrieval task of the baseline and CMCH are shown in Fig. 4. It is impressive that our method has the largest area under the curves most of the time, demonstrating that it can better balance accuracy and recall so that the CMCH method can return more relevant samples in one online query with fixed hash code length. This excellent performance is mainly due to the following reasons: (1) full exploration of complementary information from different modalities utilizing the self-adaptative multi-source feature fusion strategy instead of separately learning the different modal features. (2) Applying a novel "one-step" discrete hash optimization method avoiding quantization loss, preserves more representative information into every bit.

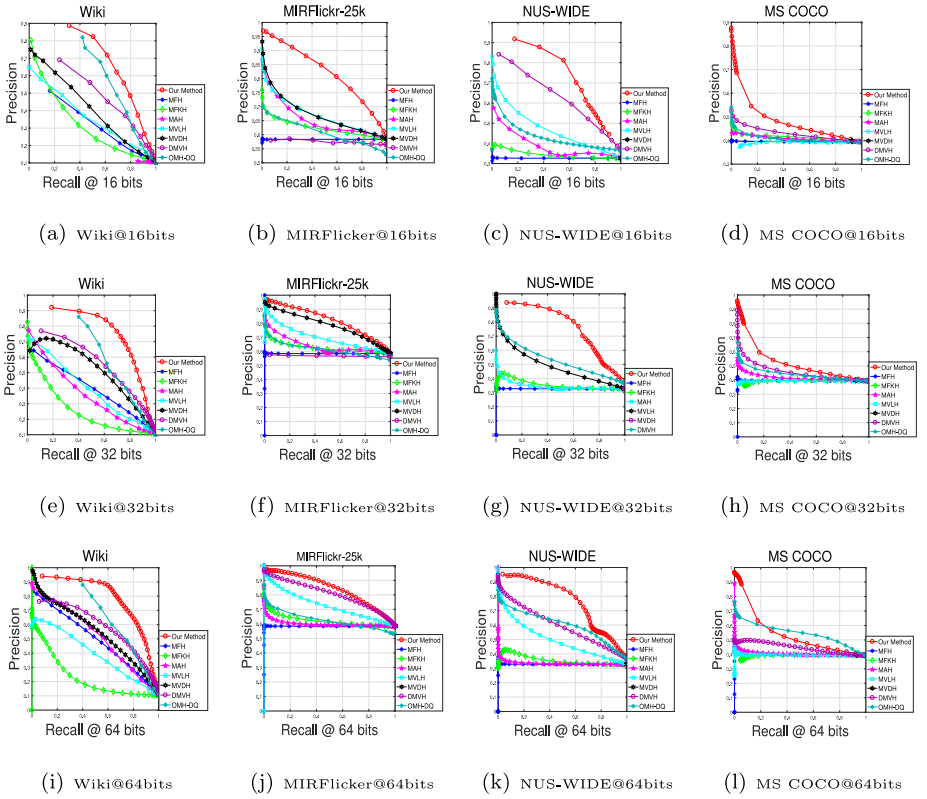


Fig. 4. Performance variations of precision–recall curves on Wiki, MIRFlicker, NUS-WIDE, and MS COCO with 16, 32, and 64 bits hash codes.

Table 3
The mAP variations with training size of CMCH with 32 bits hash codes on 4 datasets.

Wiki	350	700	1050	1400	1750	2100
mAP	0.4078	0.6459	0.7168	0.7441	0.7447	0.7471
MIRFlicker	2500	5000	7500	10,000	12,500	15,000
mAP	0.6947	0.7763	0.8091	0.8389	0.8598	0.8604
MS COCO	5000	10,000	15,000	20,000	25,000	30,000
mAP	0.5600	0.5686	0.5547	0.5675	0.5792	0.5817
NUS-WIDE	5000	10,000	15,000	20,000	25,000	30,000
mAP	0.7080	0.7456	0.7444	0.7571	0.7568	0.7590

5.2. Accuracy variations with iterations and training sizes

In this subsection, we design an experiment on the training period of our proposed CMCH method to validate its training efficiency. Specifically, we explore the variations of mAP as the training scale increases, as shown in Table 3.

(1) With the rising of training size, the mAP firstly increases and then tends to be stable. Taking Wiki, for instance, when the training volume increases from 350 to 1400, the mAP increases by 0.3463, but when the training size changes from 1400 to 2100, the mAP only increases by 0.0030.

(2) The results are same with the MIRFlicker, MS COCO, and NUS-WIDE when the training size increases from 1,2500 to 1,5000, 25,000 to 30,000, and 25,000 to 30,000, respectively. These results indicate that CMCH can achieve satisfactory performance within specific training samples. This stable property attributes to deep flexible features of specific semantic knowledge incorporated to enhance the representation of hash codes. Meanwhile, the most representative samples in the training set can be selected sequentially by the cognitive curriculum learning mode, which accelerates the learning process and contributes to the efficient learning of CMCH.

5.3. Efficiency analysis and discussion

This subsection illustrates the efficiency of CMCH by experimentally comparing its computational time with baselines. The time comparisons of training and querying periods on four benchmark datasets with the fixed 32 bit hash codes are presented in Table 4.

Table 4
The training and query time of compared methods and CMCH on 4 datasets with 32 bits hash codes.

Methods	Wiki		MIR Flickr		NUS-WIDE		MS COCO	
	Training (s)	Query (s)	Training (s)	Query (s)	Training (s)	Query (s)	Training (s)	Query (s)
MFH	2.026	**0.002**	23.01	0.052	21.11	0.059	645.1	0.261
MFKH	2.152	0.023	4.261	0.054	8.650	0.053	104.6	0.122
MAH	7.762	0.134	61.49	1.364	59.94	1.398	2985	19.10
MVLH	41.39	0.771	31.21	14.85	68.78	18.74	299.5	184.0
MVDH	1.445	0.804	32.93	20.26	64.32	27.18	228.8	223.1
DMVH	151.9	0.077	460.4	0.893	475.6	0.915	10,650	12.78
OMH-DQ	**1.316**	0.084	**4.159**	0.223	**8.502**	1.708	**37.19**	1.275
Ours	2.009	0.032	7.148	**0.049**	281.2	**0.042**	85.61	**0.116**

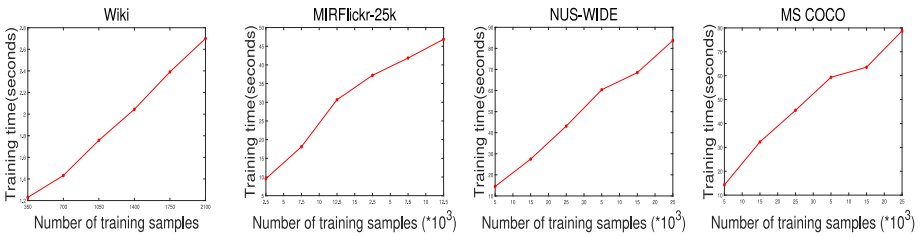


Fig. 5. The training time variations with training size on 4 different datasets of 32 bits hash codes.

We can find that the time consumption of CMCH is less than or comparable with most of the baselines both in the training and querying processes, always ranking top-three training efficiency on three datasets. Significantly, the CMCH has superior efficiency on the querying stage, indicating that even has to calculate fusion weights for every online querying, the learned hash function consumes less time than the existing alternative methods. Furthermore, as shown in Fig. 5, its training time increases linearly with the growth of the training scale. This impressive linear computational complexity of training samples further confirms that CMCH is suitable for large-scale datasets.

The high efficiency of CMCH mainly profits from these novelties: (1) The proposed one-step discrete optimization algorithm, learning hash codes through one-step discrete calculation, speeds up the resulting calculation and avoids quantization errors. (2) Collaboratively mapping multi-source features into latent unified representations in nonlinear kernel space as the hashing learning base avoids the storage of pre-construct graphs, decreasing the complexity. (3) Direct application of the original semantic label matrix as supervision avoids the redundant approximate calculation and the decomposition of the pairwise similarity matrix.

5.4. Ablation study

5.4.1. Ablation study of nonlinear-kernel collaborative multi-modal feature embedding

The proposed CMCH utilizes the RBF kernel mapping scheme to integrally map heterogeneous feature representations of different modalities into unified representation vectors while preserving the properties of original data. These vectors not only preserve more complementary semantics by reducing the architectural gaps of different modalities but also extract the nonlinear relations of samples, so they have stronger description ability than raw features. To validate the effectiveness of this nonlinear feature mapping, a variant of CMCH, named CMCH-norbf, is designed for comparison, which simply imports the original features of different modalities into the model. The comparison of mAP scores of CMCH-norbf and CMCH on the four datasets is shown in Table 5. We can find that without the nonlinear feature fusion strategy, the accuracy of CMCH-norbf dramatically declines. This result testifies the expected observation and indicates that the nonlinear mapping significantly enhances the representative performance of the learned hash codes.

5.4.2. Ablation study of self-adaptive weighing scheme

When embedding the multi-modal data into latent communal space, we propose a self-adaptive modality weighting strategy to collaboratively capture their complementarity. To evaluate its effectiveness, we design a variant of CMCH named CMCH-equ, which applies the mean weight for each modality during the training and querying processes without differentiating their contributions. The comparison of four datasets with hash codes lengths of 16bits, 32bits, 64bits, 128bits is shown in Table 5. Obviously, CMCH performs better than the variant with fixed fusion weights on all of the four datasets. With adaptively generating modal weights, our self-adaptive multi-modal fusion module can dynamically capture different modalities' complementarity, thus the self-adaptive weighing scheme is critical and indispensable to improve retrieval performance.

Table 5
The mAP of CMCH and its 8 variants on 4 datasets with 16, 32, 64, and 128 bits hash codes.

Methods	Wiki				MIRFlicker			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CMCH-equ	0.6612	0.7344	0.7559	0.7504	0.8261	0.8363	0.8606	**0.8798**
CMCH-nospl	0.6574	0.7037	0.7230	0.7338	0.8236	0.8090	0.8463	0.8705
CMCH-relax	0.7110	0.7176	0.7386	0.7433	0.8097	0.8450	0.8618	0.8680
CMCH-norbf	0.6950	0.7161	0.7241	0.7276	0.7796	0.7733	0.8171	0.8467
CMCH	**0.7449**	**0.7656**	**0.7675**	**0.7648**	**0.8303**	**0.8497**	**0.8631**	0.8728
Methods	NUS-WIDE				MS COCO			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CMCH-equ	0.7073	0.7672	0.7898	0.7976	0.5122	0.5624	0.6365	0.6655
CMCH-nospl	0.6986	0.7471	0.7809	0.7882	0.5387	0.5838	0.6386	0.6596
CMCH-relax	0.7215	0.7574	0.7780	0.7915	0.5290	0.6042	0.6398	0.6591
CMCH-norbf	0.3552	0.3519	0.3607	0.3624	0.4983	0.5111	0.5615	0.6002
CMCH	**0.7511**	**0.7681**	**0.7910**	**0.7988**	**0.5596**	**0.6144**	**0.6504**	**0.6773**

5.4.3. Ablation study of curriculum learning mode

We introduce a curriculum learning mode, *i.e.*, self-paced learning to progressively conduct feature aggregation and improve the robustness to noise of the model. To evaluate its effectiveness, a variant of CMCH, CMCH-nospl, is conducted for comparison, which randomly selects the sample sequence in the training period. The performance comparison of CMCH and CMCH-nospl method on four datasets is shown in Table 5. We can see that the representative information will be lost from the learned hash codes without the cognitive learning strategy because the 'easy' to 'complex' sample sequence structured by self-paced learning characterizes a more sensible learning path for generating more representative unified hash codes.

5.4.4. Ablation study of the proposed discrete optimization

We propose a fast one-step efficient discrete optimization mode to directly learn hash codes instead of the commonly-used two-step relaxing strategy with unnecessary quantization loss. For comparison, we design a variant of CMCH named CMCH-relax, which first relaxes the discrete constraint to learn continuous codes and then applies mean-thresholding to obtain discrete hash codes. The optimizing phase of CMCH-relax is basically similar to Eq. (18) except for the solution of $\mathbf{B}$. Specifically, the CMCH-relax obtains $\mathbf{B}$ by getting the solution to the continuity equation written as $\mathbf{B} = \frac{\beta \mathbf{R}\mathbf{V} + \theta \mathbf{P}_N \mathbf{Y}}{\beta + \theta}$ following by a thresholding strategy. The retrieval performance of CMCH-relax on four datasets is shown in Table 5. As can be seen in the promising experimental results, the proposed one-step discrete optimization effectively generates avoiding quantization loss and obtains very competing results, improving large-scale retrieval precision.

5.5. Visualization results

In this subsection, we apply the t-distributed stochastic neighbor embedding (t-SNE) to visualize the learned high-dimensional data features of CMCH in a two-dimensional map, in which similar objects are modeled into neighbors and different objects into distant points. Fig. 6 plots the comparison of features clusters learned by MFH, MAH, and our method after dimension-reduction of 32 bits hash codes on the four datasets, where each point represents a semantical feature corresponding to one multimodal sample and specific color represents a classified tag. As can be seen, the proposed CMCH outperforms competing methods by a large preponderance.

Take the single-label dataset Wiki, for example. Not only "red" but also "green" communities in MFH overlap with others, and most of the ten categories are mixed together by the MAH method. On the contrary, CMCH clearly divides all of the samples into ten clusters. For the other three multi-label datasets, MFH can barely not separate categories. Although MAH can learn some of the categories, it still ignores some labels. For example, most of the "green" is mixed with "blue" points. However, similar to the result of Wiki, our CMCH method can also accurately learn features of multi-label samples. Therefore, this visualization experiment further verifies that our CMCH can capture the latent representative features of different categories from samples enhancing the representational capacity of the learned hash codes.

In a words, we can find that the inhomogeneous multi-modal features learned by CMCH are well separated over the plane, and ones in the same congener gather better clusters than the compared methods in the latent communal space on all of the datasets, confirming that CMCH executes more accurate similarity preservation.

5.6. Parameter sensitivity analysis and discussion

There are three regularization parameters β, λ, θ in the CMCH objective function Eq. (7), and p is the number of random anchors. We utilize the commonly-used grid search method to obtain the optimal value of each parameter with fixing the others. In specific, for β, we tune it from 0.1 to 0.9 with step size 0.1, and θ the same with λ both vary from $\{10^{-5}, 10^{-3}, 10^{-1}, 1, 10, 10^{3}, 10^{5}\}$. We tune p in the range of $\{100, 300, 500, 700, 900, 1000\}$. Fig. 7 shows the detailed experimental results of parameter sensitivity on MIRFlicker with parameters of 32 bits hash codes, and the results are similar on other datasets with different lengths of binary codes. From the curves, it can be seen that CMCH performs best when β, θ, λ and p are set to $0.5, 10, 10^{-7}, 500$, respectively.

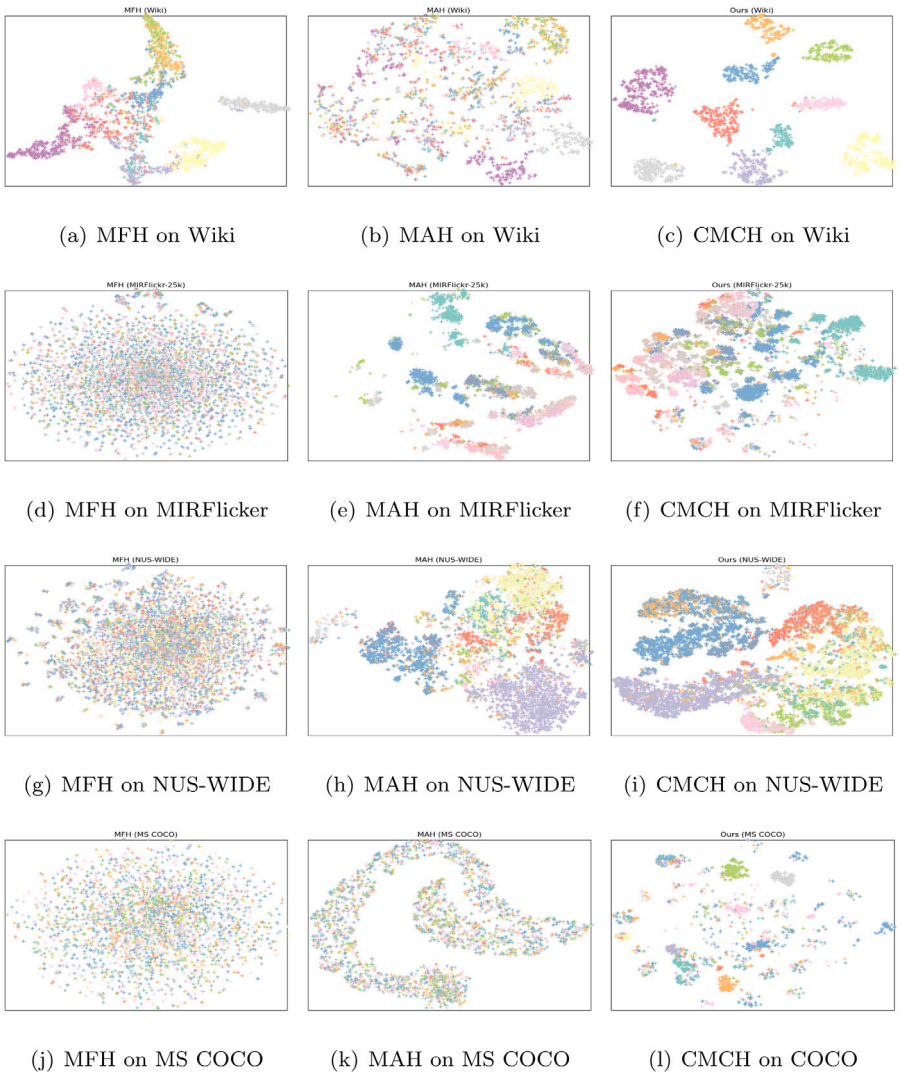


Fig. 6. T-SNE comparisons of 32 bits hash codes of MFH, MAH and CMCH on 4 datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

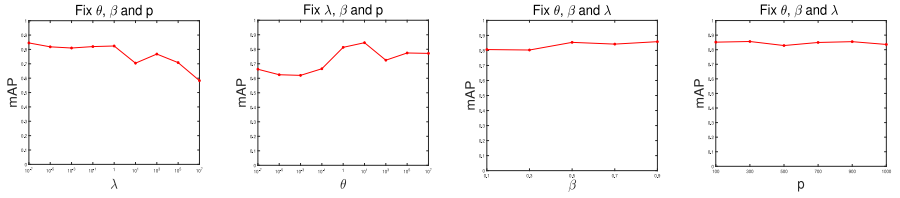
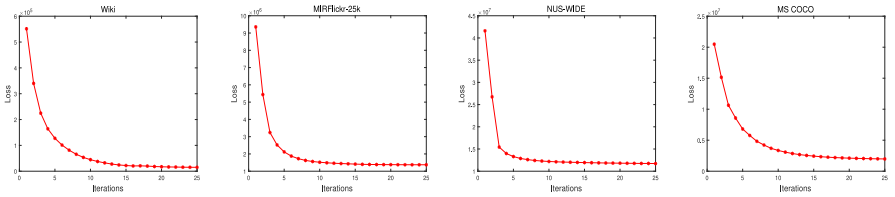


Fig. 7. mAP variations with 4 parameters on MIRFlickr with 32 bits hash codes.

5.7. Convergency analysis

Because of the mathematical definition of every term, our object function Eq. (7) is bounded to be positive. Moreover, according to Section 3.6, under the optimization steps of each variable in all iterations, the object loss function value will be monotonically decreased or equal to the previous. Therefore, according to the monotone convergence theorem, our proposed CMCH method can



(a) Loss on Wiki (b) Loss on MIRFlicker (c) Loss on NUS-WIDE (d) Loss on MS COCO

Fig. 8. The objective function values variations with iterations of 32 bits hash codes on 4 different datasets.

theoretically converge under iterative mannered optimization. Using the parameters obtained in the previous step, we construct convergence experiments to visually evaluate the convergent nature of CMCH as shown in Fig. 8. We can observe that the value of the loss function on each dataset stably decreases with iteration number increasing and eventually converges within 15 iterations. The primary reason of this fast convergence property expectantly owes to the analytical or optimal solutions to subproblems.

6. Conclusion

In this work, we proposed a novel efficient discriminative multimodal GIR hash learning framework yielding state-of-the-art multimodal retrieval performance, dubbed cognitive multi-modal consistent hashing with flexible semantic transformation (CMCH), which could intelligently construct the cooperative latent space with heterogeneous multisource data supervised by a flexible semantic transformation strategy. Specifically, we formulated an adaptive modality fusion manner to seamlessly coalesce the modal complementarity, which could improve the online retrieval precision with the dynamically constructed informative objective hash codes for different queries. In addition, a cognitive self-paced sample selecting strategy is introduced to autonomously construct the learning sequence, significantly increasing the model robustness and efficiency. Furthermore, we leveraged a specific pointwise category label matrix with a deep transform learning module that flexibly distilled discriminative semantic knowledge and interactively guided the objective hash code learning, readily settling the pre-constructing large similar graph and generating more representative hash codes. Extensive empirical studies of large-scale multi-modal retrieval on four geo-media relevant public social datasets validated that the proposed CMCH method could achieve superior performance than the state-of-the-art multimodal hashing methods in terms of standard evaluation metrics.

CRediT authorship contribution statement

Junfeng An: Conceptualization, Experiments, Writing. **Haoyang Luo:** Experiments, Methodology. **Zheng Zhang:** Writing – original draft, Methodology, Data collection, Supervision, Review, Revision. **Lei Zhu:** Data Collection, Methodology. **Guangming Lu:** Supervision, Revision, Visualization, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grants Nos. 62002085, 62176077), the Guangdong Basic and Applied Basic Research Foundation (Grants Nos. 2019A1515110475, 2019Bl515120055), and also supported in part by Special Projects for Key Fields in Higher Education of Guangdong, China (2021ZDZX1042).

References

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al. (2007). Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, *19*, 153.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41–48).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cao, W., Feng, W., Lin, Q., Cao, G., & He, Z. (2020). A review of hashing methods for multimodal retrieval. *IEEE Access*, *8*, 15377–15391.

Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). 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* (pp. 1–9).

Dang-Nguyen, D. T., Boato, G., Moschitti, A., & Natale, F. (2012). Supervised models for multimodal image retrieval based on visual, semantic and geographic information. In *CBMI*.

Gong, Y., Lazebnik, S., Gordo, A., & Perronnin, F. (2013). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2916–2929. http://dx.doi.org/10.1109/TPAMI.2012.193.

Guo, J., Chang, H., & Zhu, W. (2020). Preserving ordinal consensus: Towards feature selection for unlabeled data. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 34* (pp. 75–82).

He, J., Du, C., Zhuang, F., Yin, X., & Long, G. (2020). Online Bayesian max-margin subspace learning for multi-view classification and regression. *Machine Learning*, *109*(6).

Huiskes, M. J., & Lew, M. S. (2008). The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on multimedia information retrieval* (pp. 39–43).

Kang, Y., Kim, S., & Choi, S. (2012). Deep learning to hash with multiple representations. In *2012 IEEE 12th international conference on data mining* (pp. 930–935). http://dx.doi.org/10.1109/ICDM.2012.24.

Kim, S., & Choi, S. (2013). Multi-view anchor graph hashing. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3123–3127). IEEE.

Kumar, C., Heuten, W., & Boll, S. (2013). Visualization support for multi-criteria decision making in geographic information retrieval. In *Availability, reliability, and security in information systems and hci*.

Li, Q., Sun, Z., He, R., & Tan, T. (2017). Deep supervised discrete hashing. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 2479–2488).

Li, W. J., Wang, S., & Kang, W. C. (2016). Feature learning based deep supervised hashing with pairwise labels. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 1711–1717).

Liang, X., Shen, J., Han, J., Lei, Z., & Ling, S. (2017). Dynamic multi-view hashing for online image retrieval. In *The 26th international joint conference on artificial intelligence*.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Lin, G., Shen, C., Suter, D., & Van Den Hengel, A. (2013). A general two-step approach to learning-based hashing. In *Proceedings of the IEEE international conference on computer vision* (pp. 2552–2559).

Liu, X., He, J., Liu, D., & Lang, B. (2012). Compact kernel hashing with multiple features. In *Proceedings of the 20th ACM international conference on multimedia* (pp. 881–884).

Liu, W., Wang, J., Ji, R., Jiang, Y. G., & Chang, S. F. (2012). Supervised hashing with kernels. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2074–2081).

Liu, W., Wang, J., Kumar, S., & Chang, S. F. (2011). Hashing with graphs. In *Proceedings of international conference on machine learning* (pp. 1–8).

Liu, L., Yu, M., & Shao, L. (2015). Multiview alignment hashing for efficient image search. *IEEE Transactions on Image Processing*, *24*(3), 956–966.

Liu, L., Zhang, Z., & Huang, Z. (2020). Flexible discrete multi-view hashing with collective latent feature learning. *Neural Processing Letters*, *52*(3), 1765–1791.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision, Vol. 2* (pp. 1150–1157). http://dx.doi.org/10.1109/ICCV.1999.790410.

Lu, X., Zhu, L., Cheng, Z., Nie, L., & Zhang, H. (2019). Online multi-modal hashing with dynamic query-adaption. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 715–724).

Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. *Advances in Neural Information Processing Systems*, *23*, 1813–1821.

Pereira, J. C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G. R., Levy, R., et al. (2013). On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(3), 521–535.

Purificato, E., & Rinaldi, A. M. (2018). Multimedia and geographic data integration for cultural heritage information retrieval. *Multimedia Tools and Applications*, *77*(20), 27447–27469.

Raginsky, M., & Lazebnik, S. (2009). Locality-sensitive binary codes from shift-invariant kernels. *Advances in Neural Information Processing Systems*, *22*, 1509–1517.

Ravishankar, S., Wen, B., & Bresler, Y. (2015). Online sparsifying transform learning— part I: Algorithms. *IEEE Journal of Selected Topics in Signal Processing*, *9*(4), 625–636.

Shen, X., Shen, F., Liu, L., Yuan, Y. H., Liu, W., & Sun, Q. S. (2018). Multiview discrete hashing for scalable multimedia search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *9*(5), 1–21.

Shen, X., Shen, F., Sun, Q. S., & Yuan, Y. H. (2015). Multi-view latent hashing for efficient multimedia search. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 831–834). Association for Computing Machinery.

Shen, F., Xu, Y., Liu, L., Yang, Y., Huang, Z., & Shen, H. T. (2018). Unsupervised deep hashing with similarity-adaptive and discrete optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 3034–3044.

Shen, H. T., Zhu, X., Zhang, Z., Wang, S.-H., Chen, Y., Xu, X., et al. (2021). Heterogeneous data fusion for predicting mild cognitive impairment conversion. *Information Fusion*, *66*, 54–63.

Song, J., Yang, Y., Huang, Z., Shen, H. T., & Luo, J. (2013). Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia*, *15*(8), 1997–2008. http://dx.doi.org/10.1109/TMM.2013.2271746.

Wang, Y., Luo, X., & Xu, X. S. (2020). Label embedding online hashing for cross-modal retrieval. In *The 28th ACM international conference on multimedia*.

Wang, Z., Zhang, Z., Luo, Y., Huang, Z., & Shen, H. T. (2020). Deep collaborative discrete hashing with semantic-invariant structure construction. *IEEE Transactions on Multimedia*, http://dx.doi.org/10.1109/TMM.2020.2995267.

Wang, J., Zhang, T., song, j., Sebe, N., & Shen, H. T. (2018). A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 769–790. http://dx.doi.org/10.1109/TPAMI.2017.2699960.

Weiss, Y., Torralba, A., Fergus, R., et al. (2008). Spectral hashing. In *Neural information processing systems, Vol. 1* (p. 4). Citeseer.

Xu, C., Tao, D., & Xu, C. (2015). Multi-view self-paced learning for clustering. In *Proceedings of the 24th international conference on artificial intelligence* (pp. 3974–3980).

Yang, R., Shi, Y., & Xu, X. S. (2017). Discrete multi-view hashing for effective image retrieval. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval* (pp. 175–183). http://dx.doi.org/10.1145/3078971.3078981.

Zhang, Z., Lai, Z., Huang, Z., Wong, W. K., Xie, G., Liu, L., et al. (2019). Scalable supervised asymmetric hashing with semantic and latent factor embedding. *IEEE Transactions on Image Processing*, *28*(10), 4803–4818.

Zhang, Z., Liu, L., Qin, J., Zhu, F., Shen, F., Xu, Y., et al. (2018). Highly-economized multi-view binary compression for scalable image clustering. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 717–732).

Zhang, Z., Liu, L., Shen, F., Shen, H. T., & Shao, L. (2019). Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(7), 1774–1782.

Zhang, D., Wang, F., & Si, L. (2011). Composite hashing with multiple information sources. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 225–234).

Zheng, C., Zhu, L., Lu, X., Li, J., Cheng, Z., & Zhang, H. (2019). Fast discrete collaborative multi-modal hashing for large-scale multimedia retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *32*(11), 2171–2184.

Zhu, X., Li, H., Shen, H. T., Zhang, Z., Ji, Y., & Fan, Y. (2021). Fusing functional connectivity with network nodal information for sparse network pattern learning of functional brain networks. *Information Fusion*, *75*, 131–139.

Zhu, L., Lu, X., Cheng, Z., Li, J., & Zhang, H. (2020). Flexible multi-modal hashing for scalable multimedia retrieval. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(2), 1–20.

Zhu, L., Shen, J., Xie, L., & Cheng, Z. (2017). Unsupervised visual hashing with semantic assistant for content-based image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *29*(2), 472–486. http://dx.doi.org/10.1109/TKDE.2016.2562624.