Unsupervised Dual Deep Hashing With Semantic-Index and Content-Code for Cross-Modal Retrieval

Bin Zhang, Yue Zhang, Junyu Li, Jiazhou Chen, *Member, IEEE*, Tatsuya Akutsu, *Senior Member, IEEE*, Yiu-Ming Cheung, *Fellow, IEEE*, and Hongmin Cai, *Senior Member, IEEE*

***Abstract*—Hashing technology has exhibited great cross-modal retrieval potential due to its appealing retrieval efficiency and storage effectiveness. Most current supervised cross-modal retrieval methods heavily rely on accurate semantic supervision, which is intractable for annotations with ever-growing sample sizes. By comparison, the existing unsupervised methods rely on accurate sample similarity preservation strategies with intensive computational costs to compensate for the lack of semantic guidance, which causes these methods to lose the power to bridge the semantic gap. Furthermore, both kinds of approaches need to search for the nearest samples among all samples in a large search space, whose process is laborious. To address these issues, this paper proposes an unsupervised dual deep hashing (UDDH) method with semantic-index and content-code for cross-modal retrieval. Deep hashing networks are utilized to extract deep features and jointly encode the dual hashing codes in a collaborative manner with a common semantic index and modality content codes to simultaneously bridge the semantic and heterogeneous gaps for cross-modal retrieval. The dual deep hashing architecture, comprising the head code on semantic index and tail codes on modality content, enhances the efficiency for cross-modal retrieval. A query sample only needs to search for the retrieved samples with the same semantic index, thus greatly shrinking the search space and achieving superior retrieval efficiency. UDDH integrates the learning processes of deep feature extraction, binary optimization, common semantic index, and modality content code within a unified model, allowing for collaborative optimization to enhance the overall performance. Extensive experiments are conducted to demonstrate the retrieval superiority of the proposed approach over the state-of-the-art baselines.**

***Index Terms*—Binary optimization, cross-modal retrieval, deep hashing, dual coding, retrieval of similar content, sample assignment, semantic index, unsupervised learning.**

I. Introduction

Recently, with the rapid growth of massive amount of multimodal data accumulated from various modalities, cross-modal retrieval has attracted considerable attention. This task aims to search for the nearest samples in the reference database from one modality with the smallest distance to the query sample from another modality. This nearest neighbor search strategy incurs a very large computational burden and experiences the curse of dimensionality when facing large-scale high-dimensional multimodal data [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. Hashing technology has been recognized as an effective solution for facilitating fast cross-modal retrieval due to its appealing retrieval efficiency [15], [16], [17], [18], [19], [20], [21], [22], [23]. It utilizes compact binary codes to represent high-dimensional data with a low-complexity nearest-neighbor search process in the Hamming space. The traditional supervised cross-modal retrieval methods transform multimodal features into a common semantic space to bridge the semantic gap, where the multimodal samples with the same semantic label are expected to be encoded as a unified hashing code [24], [25], [26], [27], [28]. For example, matrix tri-factorization hashing (MTFH) [28] decomposes the semantic correlation matrix calculated on the given semantic labels into multimodal hashing codes with different lengths for paired or unpaired multimodal data. These methods make the learning processes of feature representation and binary optimization completely independent, which may therefore result in suboptimal solutions. Recently, with the powerful feature representation abilities, deep hashing-based cross-modal retrieval methods have utilized multiple deep neural networks to simultaneously extract deep multimodal features and encode them as binary codes [29]. These binary codes are optimally compatible with deep feature representations, thus yielding discriminative

Received 9 December 2022; revised 3 March 2024; accepted 16 September 2024. Date of publication 24 September 2024; date of current version 4 December 2024. This work was supported in part by the Key-Area Research and Development Program of Guangzhou City under Grant 202206030009, Grant 2023B01J0022, and Grant 2023B01J1001, in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2021B0909060002, in part by the Key-Area of Guangdong Provincial Department of Education under Grant 2021ZDZX1005, in part by the National Key Research and Development Program of China under Grant 2022YFE0112200, in part by the National Natural Science Foundation of China under Grant U21A20520, Grant 62325204, Grant 62172112, Grant 62102153, and Grant 62402123, and in part by the China Postdoctoral Science Foundation under Grant 2023M730752. Recommended for acceptance by B. Han. *(Corresponding author: Hongmin Cai.)*

Bin Zhang is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Guangdong Institute of Intelligence Science and Technology, Zhuhai, Guangdong 519031, China.

Yue Zhang is with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China.

Junyu Li, Jiazhou Chen, and Hongmin Cai are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: hmcai@scut.edu.cn).

Tatsuya Akutsu is with the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan.

Yiu-Ming Cheung is with the Department of Computer Science, Hong Kong Baptist University Hong Kong, SAR, China.

Digital Object Identifier 10.1109/TPAMI.2024.3467130

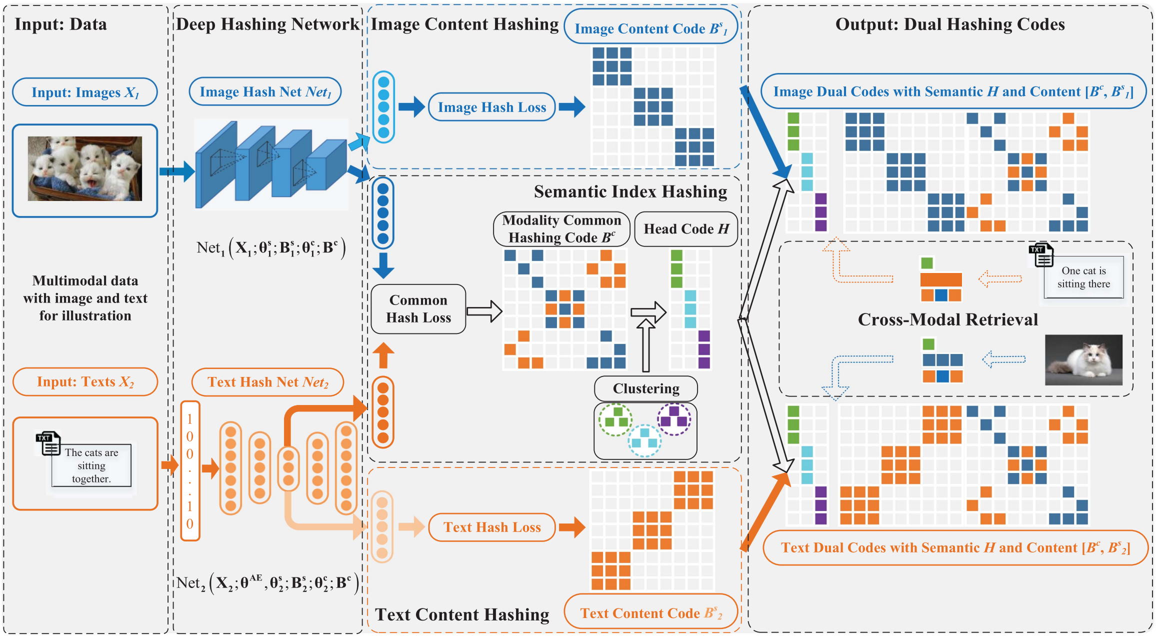

Input: Data
Input: Images X_1
Multimodal data with image and text for illustration
Input: Texts X_2
The cats are sitting together.
Deep Hashing Network
Image Hash Net Net_1
$\mathrm{Net}_1\left(\mathbf{X}_1;\theta_1^s;\mathbf{B}_1^s;\theta_1^c;\mathbf{B}^c\right)$
Text Hash Net Net_2
1 0 0 . . . 1 0
$\mathrm{Net}_2\left(\mathbf{X}_2;\theta^{AE},\theta_2^s;\mathbf{B}_2^s;\theta_2^c;\mathbf{B}^c\right)$
Image Content Hashing
Image Content Code $B^s{}_1$
Image Hash Loss
Semantic Index Hashing
Modality Common Hashing Code B^c
Head Code H
Common Hash Loss
Clustering
Text Hash Loss
Text Content Hashing
Text Content Code $B^s{}_2$
Output: Dual Hashing Codes
Image Dual Codes with Semantic H and Content $[B^c, B^s{}_1]$
One cat is sitting there
Cross-Modal Retrieval
Text Dual Codes with Semantic H and Content $[B^c, B^s{}_2]$

Fig. 1. The framework of unsupervised dual deep hashing (UDDH) with semantic-index and content-code for cross-modal retrieval. In detail, we show the cross-modal retrieval process with two modalities of image and text for illustration purposes, with the modules of deep hashing network, semantic index hashing and multimodal content hashing of image and text. Subsequently, the multimodal data are coded as dual hashing codes with head code on the common semantic index $\boldsymbol{H}$, and tail codes on the modality content $[\boldsymbol{B^c}, \boldsymbol{B_i^s}]$ consisting of modality-shared content code $\boldsymbol{B^c}$ and multimodal content code $\boldsymbol{B_i^s}$. For both the head and tail code matrices, a row represents a sample, while a column represents a feature. To simplify the binary representation, we use white squares to represent -1 or 0, while the other colors represent 1.

features and high-quality hashing codes. Consequently, supervised deep hashing methods have achieved great success and superior retrieval performance under semantic supervision [6], [8], [30], [31], [32], [33], [34]. For example, deep semantic-preserving ordinal hashing (DSPOH) [34] utilizes deep neural networks to learn hash functions while preserving feature rankings through ordinal embedding. DSPOH is expected to jointly accomplish compatible hashing function learning and semantic label prediction. However, such a supervised strategy requires laborious sample labeling which is intractable in real-world applications.

Alternatively, unsupervised deep hashing methods have recently been explored to avoid the manual annotation procedure [35], [36], [36], [37], [38], [39], [40], [41], [42], [43]. These techniques usually map the multimodal feature space into a unified Hamming space to bridge the heterogeneous gap while preserving the sample similarity in both spaces. For example, unsupervised deep fusion cross-modal hashing (UDFCH) [39] learns a unified hashing code across all modalities by preserving the intramodality and intermodality similarities through variational inference.

However, unsupervised methods rely heavily on accurate sample similarity to compensate for the lack of semantic guidance when encoding multimodal data with a unified hashing code. They fail to exploit the common semantic information using only the sample similarity and thus lose the power to bridge the semantic gap, which is significant for cross-modal retrieval. Additionally, computing the sample similarity is expensive, making it inapplicable to real-world retrieval applications with ever-growing sample sizes. Therefore, it poses a challenge in acquiring deep hashing codes for cross-modal retrieval, with efficiently retaining both united semantic information and modality content information, while simultaneously shrinking the search space.

Herein, this paper proposes an unsupervised dual deep hashing (UDDH) method with semantic-index and content-code for cross-modal retrieval, to simultaneously exploit the common semantic information and modality content information for bridging the semantic and heterogeneous gaps of multimodal data. The framework of UDDH is illustrated in Fig. 1. The UDDH method encodes each multimodal sample using dual binary codes, comprised of both head and tail codes. The head code is dedicated to preserve the common semantic index, while the tail code represents the modality-shared and multimodal content information. To be specific, UDDH employs deep hashing networks to extract multimodal features and jointly encode the dual binary codes, thus yielding discriminative features and high-quality hashing codes. The head code learning utilizes modality-shared content code with clustering assignment embedding to learn a common semantic index and modality-shared content information. The tail code learning utilizes multimodal content codes and employing a fast cross-modal affinity preservation strategy to exploit multimodal content information. These two learning schemes are incorporated seamlessly into a deep hashing model to simultaneously bridge the semantic and heterogeneous gaps for cross-modal retrieval. At last, the model could focus on the cross-modal retrieval of similar content under the guidance of the common semantic index. UDDH integrates the learning processes of deep feature extraction, binary optimization, common semantic index and modality content code within a unified model, allowing for collaborative optimization to enhance the overall performance. In the retrieval stage, this dual deep hashing design with head code for a common semantic index and tail codes for modality content makes it very efficient for cross-modal retrieval. Therefore, a query sample only needs to search for the retrieved samples with the same semantic index, thus achieving superior retrieval efficiency.

The proposed UDDH is designed to realize unsupervised cross-modal retrieval with dual deep hashing on a common semantic index and modality content code. The major contributions of the proposed work are summarized as follows:

1) The dual hashing architecture significantly enhances the efficiency for cross-modal retrieval tasks. This approach matches a query sample by retrieving samples with shared semantic index, thus greatly shrinking the search space and achieving superior retrieval efficiency.
2) The head code learning utilizes modality-shared codes incorporating with clustering assignment embedding to learn the common semantic information, with bridging the semantic gap for cross-modal retrieval.
3) The tail code learning utilizes multimodal content codes cooperating with a fast cross-modal affinity preservation strategy to exploit the multimodal content information, with bridging the heterogeneous gaps for cross-modal retrieval.
4) UDDH employs deep hashing networks to extract multimodal features and subsequently encodes them as binary codes in a collaborative manner. It integrates the learning processes of deep feature extraction, binary optimization, common semantic index, and modality content code within a unified model, allowing for collaborative optimization to enhance the overall performance.

The rest of this paper is organized as follows. We make an overview of the related work on hashing-based cross-modal retrieval in Section II. Section III provides the details of our proposed UDDH method and the corresponding numerical scheme for solving the optimization problem. Extensive experiments are conducted to investigate the performance of UDDH in comparison with the state-of-the-art baselines in Section IV. Finally, Section V draws a conclusion.

II. Overview of Related Works

Hashing technology has been recognized as an effective solution for facilitating fast cross-modal retrieval due to its appealing retrieval efficiency. It utilizes compact binary codes to represent high-dimensional data with a low complexity nearest-neighbor search process in the Hamming space. In this section, we provide a brief overview of the four categories of hashing-based cross-modal retrieval approaches: supervised hashing, unsupervised hashing, supervised deep hashing and unsupervised deep hashing.

Utilizing semantic labels, traditional supervised hashing-based cross-modal retrieval methods transform multimodal features into a shared semantic space. Consequently, samples with the same semantic label are encoded as the unified hashing code [24], [25], [26], [27], [28]. For example, matrix tri-factorization hashing (MTFH) [28] decomposes the semantic correlation matrix calculated on the given semantic labels into multimodal hashing codes with different lengths for paired or unpaired multimodal data. Fast supervised discrete hashing (FSDH) [27] regresses the multi-modal features to their semantic labels to learn the corresponding hashing codes with a one-step way instead of iterative optimization. Label consistent matrix factorization hashing (LCMFH) [24] transforms multimodal features into the semantic space where the multimodal data with the same semantic label are coded as a consistent hashing code to further exploit the label information. Although supervised methods can achieve promising retrieval performance, they rely heavily on labels produced by manual annotation, which is time-consuming and unsuitable for real-world applications.

Unsupervised hashing-based methods for cross-modal retrieval have emerged to bridge the heterogeneous gap among multimodal data without the need for semantic labels [15], [16], [17], [18], [19], [20], [21], [22], [23]. Most of these methods seek to preserve samplewise similarity when guiding the hashing code learning process. For example, joint and individual matrix factorization hashing (JIMFH) [15] learns the individual and unified hashing codes for multimodal data to preserve their specific and common properties, respectively, via individual and joint matrix factorization. Cross-modal discrete hashing (CMDH) [20] utilizes linear and nonlinear frameworks to map multimodal features into the Hamming space by learning a set of shared hashing codes to bridge the modality gap. Collective reconstructive embedding (CRE) [23] employs different modality-specific models for multimodal data to simultaneously exploit the specific and integrated information while preserving the similarities among different modalities. However, these unsupervised retrieval methods decompose hashing learning into two independent steps, namely feature representation and binary optimization, which may therefore result in suboptimal solutions.

Recently, deep hashing-based cross-modal retrieval methods have gained attention for their remarkably powerful feature representation potential. These methods leverage multiple deep neural networks to extract deep multimodal features and encode them as binary codes simultaneously [29]. The binary codes are optimally compatible with deep feature representations, thus yielding discriminative features and high-quality hashing codes.

Supervised deep hashing methods have achieved great success and superior retrieval performance under semantic supervision [6], [8], [30], [31], [32], [33], [34]. For example, deep visual-semantic hashing (DVSH) [33] consists of two modality-specific networks and a semantic-visual fusion network to jointly realize hash code learning and embedding learning. Deep semantic-preserving ordinal hashing (DSPOH) [34] utilizes deep neural networks to learn hash functions while preserving the rankings of feature through ordinal embedding. DSPOH is expected to accomplish compatible hashing function learning and semantic label prediction processes. Hierarchical semantic structure preserving hashing (HSSPH) [44] aims at learning more discriminative hashing codes by exploiting the observed label hierarchy information under the guidance of both semantic structures and labels for hierarchical semantic

TABLE I
BRIEF DESCRIPTIONS OF THE MATHEMATICAL NOMENCLATURE IN THIS PAPER

Notation	Remark
$\boldsymbol{X}_i$	Multimodal data
$\boldsymbol{C}$	Learned clustering center for K-means
$\boldsymbol{H}$	Head code of the semantic index
$\boldsymbol{B}^c$	Modality-shared content code
$\boldsymbol{B}_i^s$	Multimodal content code for modality i
$[\boldsymbol{B}^c, \boldsymbol{B}_i^s]$	Tail codes of the modality content for modality i
$\boldsymbol{L}_i$	Graph Laplacian matrix for modality i
$\boldsymbol{A}_i$	Affinity matrix for modality i
$\boldsymbol{Net}_i$	Integrated deep hashing network for modality i
$\mathbf{N}_i^s\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^s\right)$	Multimodal deep hashing network with a set of parameters $\boldsymbol{\theta}_i^s$ for modality i
$\mathbf{N}_i^c\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^c\right)$	Modality-shared deep hashing network with a set of parameters $\boldsymbol{\theta}_i^c$ for modality i
$\hat{\mathbf{N}}_i^s\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^s\right)$	The network output of $\mathbf{N}_i^s\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^s\right)$
$\hat{\mathbf{N}}_i^c\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^c\right)$	The network output of $\mathbf{N}_i^c\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^c\right)$
$\bar{\mathbf{N}}_i^s\left(x_i^p; \bar{\boldsymbol{\theta}}_i^s\right)$	Deep feature representation for the sample p within modality i
$\boldsymbol{1}$	The column vector with all elements equal to 1.

preservation. However, such a supervised strategy requires a laborious sample labeling step, which is intractable in real-world applications.

Alternatively, unsupervised deep hashing methods have been investigated to mitigate the need for manual annotation [35], [36], [36], [37], [38], [39], [40], [41], [42], [43]. These methods usually map the multimodal feature space into a unified Hamming space to bridge the heterogeneous gap while preserving the sample similarity in both spaces. For example, deep binary reconstruction (DBRC) [38] learns the correlated hashing codes for all modalities by preserving their intramodality and intermodality consistencies. By incorporating a scalable $tanh$ activation function, it is possible to simultaneously learn the deep feature representations and hashing codes via the standard backpropagation algorithm. Unsupervised deep fusion cross-modal hashing (UDFCH) [39] learns a unified hashing code across all modalities by preserving the intramodality and intermodality similarities through variational inference. Attention-guided semantic hashing (AGSH) [45] uses a novel deep network with an attention fusion scheme to efficiently encode the relevant multimodal features for cross-modal retrieval. Unsupervised generative adversarial cross-modal hashing (UGACH) [40] aims to make full use of the unsupervised feature representation ability of a generative adversarial network to exploit the underlying manifold structure of multimodal data for cross-modal retrieval. Deep joint-semantics reconstructing hashing (DJSRH) [41] integrates multimodal neighborhood information to construct a novel affinity matrix with joint semantic information for cross-modal retrieval. Joint-modal distribution-based similarity hashing (JDSH) [36] proposes a distribution-based affinity decision making and weighting strategy to construct a novel joint-modal affinity matrix with preserving the cross-modal correlations among multimodal data. Deep unsupervised cross-modal contrastive hashing (DUCH) [42] integrates contrastive, adversarial and binarization objectives into a joint model for multimodal similarity preservation, consistent cross-modal representation and representative hashing coding, respectively. Deep graph-neighbor coherence preserving network (DGCPN) [43] integrates the coexistent similarity, graph-neighbor coherence, and intermodality and intramodality consistency to preserve the comprehensive similarity of multimodal data for cross-modal retrieval.

The existing unsupervised methods rely on accurate sample similarity to compensate for the lack of semantic guidance, thereby requiring intensive computational costs and lose the power to bridge the semantic gap. Furthermore, all of the hashing-based cross-modal retrieval methods need to search for the nearest samples among all the samples in a massive search space, whose process is laborious.

III. Methodology

In this section, we provide a comprehensive explanation of our proposed method, namely UDDH, which is designed for unsupervised cross-modal retrieval. The workflow is illustrated in Fig. 1. The nomenclature used in the paper is summarized in Table I.

A. *Unsupervised Dual Deep Hashing (UDDH) With Semantic-Index and Content-Code for Cross-Modal Retrieval*

Mathematically, suppose that a multimodal dataset $\{\boldsymbol{X}_i\}_{i=1}^m$ is composed of m modalities and that the i-th modality is denoted by $\{x_{ij}\}_{j=1}^n$, where n is the number of sample. The proposed UDDH method consists of three modules, namely, deep hashing network, semantic index hashing and multimodal content hashing. The module of deep hashing network aims to extract multimodal features from the input multimodal data and encode them with compact hashing codes. The module of semantic index hashing aims to learn the modality-shared content code $\boldsymbol{B}^c$ and the semantic index $\boldsymbol{H}$ of the multimodal data. The module of multimodal content hashing aims to learn the multimodal content code $\boldsymbol{B}_i^s$ with a fast cross-modal affinity preservation strategy to maintain the cross-modal correlated content. These three modules are optimized collaboratively within a unified deep Hamming space and can be used to enhance each other.

The first module of deep hashing network aims to employ deep networks to extract the multimodal features and transform the high-dimensional multimodal features into low-dimensional hashing codes [36], [40], [41], [42], [43], this process is defined as

$$\boldsymbol{B}_i^s = sign\left(\hat{\mathbf{N}}_i^s\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^s\right)\right) \tag{1}$$

$$\boldsymbol{B}^c = sign\left(\hat{\mathbf{N}}_i^c\left(\boldsymbol{X}_i; \boldsymbol{\theta}_i^c\right)\right) \tag{2}$$

where $\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{s}} \in \{-1,1\}^{l^s \times n}$ and $\boldsymbol{B}^{\boldsymbol{c}} \in \{-1,1\}^{l^c \times n}$ are the multimodal hashing code for modality i and the modality-shared content code, respectively, with l^s and l^c being the hashing code length. $\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{s}}(\cdot)$ and $\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{c}}(\cdot)$ are the network outputs of the multimodal hashing networks $\mathrm{N}_i^s(\cdot)$ and the modality-shared hashing network $\mathbf{N}_{\boldsymbol{i}}^{\boldsymbol{c}}(\cdot)$, where $\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{s}}$ and $\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{c}}$ are their network weights, respectively. The multimodal hashing networks $\mathbf{N}_{\boldsymbol{i}}^{\boldsymbol{s}}(\cdot)$ and the modality-shared hashing network $\mathbf{N}_{\boldsymbol{i}}^{\boldsymbol{c}}(\cdot)$ aim to extract multimodal features from the input multimodal data $\boldsymbol{X}_{\boldsymbol{i}}$ and then encode them with the multimodal hashing code $\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{s}}$ for modality i and the modality-shared content code $\boldsymbol{B}^{\boldsymbol{c}}$, respectively. Here, $sign(\cdot)$ indicates the sign operation.

To exemplify the prototype, we use the integration of text and images as a demonstration. Specifically, we utilize the deep ResNet18 [46] network pretrained on the ImageNet database [47] to learn the representations for the image modality. Different from ResNet18, the top fully connected layer and classification layer are replaced with a new fully connected layer and a regression layer to measure the discrepancies between the network outputs $\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{s}}(\cdot)$, $\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{c}}(\cdot)$ and the hashing codes $\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{s}}$, $\boldsymbol{B}^{\boldsymbol{c}}$ (as illustrated in Fig. 1) [36], [40], [41], [42], [43]. The integrated image hashing network $\mathbf{Net_1}(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{s}}; \boldsymbol{B}_{\mathbf{1}}^{\boldsymbol{s}}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{c}}; \boldsymbol{B}^{\boldsymbol{c}})$ consisting of an image modality network $\boldsymbol{N}_{\mathbf{1}}^{\boldsymbol{s}}(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{s}})$ and a modality-shared network $\boldsymbol{N}_{\mathbf{1}}^{\boldsymbol{c}}(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{c}})$ is formulated as:

$$\begin{aligned} &\min_{\boldsymbol{\theta}_1^s, \boldsymbol{\theta}_1^c, \boldsymbol{B}_1^s, \boldsymbol{B}^c} \mathbf{Net_1}\left(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{s}}; \boldsymbol{B}_{\mathbf{1}}^{\boldsymbol{s}}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{c}}; \boldsymbol{B}^{\boldsymbol{c}}\right) \\ &= \min_{\boldsymbol{\theta}_1^s, \boldsymbol{\theta}_1^c, \boldsymbol{B}_1^s, \boldsymbol{B}^c} \begin{pmatrix} ||\hat{\boldsymbol{N}}_{\mathbf{1}}^{\boldsymbol{s}}\left(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{s}}\right) - \boldsymbol{B}_{\mathbf{1}}^{\boldsymbol{s}}||^2 \\ +||\hat{\boldsymbol{N}}_{\mathbf{1}}^{\boldsymbol{c}}\left(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{c}}\right) - \boldsymbol{B}^{\boldsymbol{c}}||^2 \end{pmatrix} \\ &\text{s.t.} \quad \boldsymbol{B}_{\mathbf{1}}^{\boldsymbol{s}} \in \{-1,1\}^{l^s \times n}, \boldsymbol{B}^{\boldsymbol{c}} \in \{-1,1\}^{l^c \times n} \end{aligned} \tag{3}$$

where the optimization of sign function $sign(\cdot)$ is relaxed with the nonlinear activation function $tanh(\cdot)$. $||\hat{\mathbf{N}}_{\mathbf{1}}^{\boldsymbol{s}}(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{s}}) - \boldsymbol{B}_{\mathbf{1}}^{\boldsymbol{s}}||^2$ and $||\hat{\mathbf{N}}_{\mathbf{1}}^{\boldsymbol{c}}(\boldsymbol{X_1}; \boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{c}}) - \boldsymbol{B}^{\boldsymbol{c}}||^2$ indicate the image hashing loss and modality-shared hashing loss, which aim to extract visual features from the input images $\boldsymbol{X_1}$ and then encode them with the image modality content codes $\boldsymbol{B}_{\mathbf{1}}^{\boldsymbol{s}}$ and modality-shared content code $\boldsymbol{B}^{\boldsymbol{c}}$, with $\hat{\mathbf{N}}_{\mathbf{1}}^{\boldsymbol{s}}(\cdot)$ and $\hat{\mathbf{N}}_{\mathbf{1}}^{\boldsymbol{c}}(\cdot)$ being the network outputs.

For the text modality with topic vectors or textual tag occurrence vectors, first, we utilize a deep autoencoder to learn discriminative feature representation. Then it cooperates with text hashing network which consists of multiple fully connected layers to learn high-quality hashing codes (as illustrated in Fig. 1). It is worth noting that the encoder parameters are shared between both the deep autoencoder and text hashing networks. Consequently, the deep autoencoder is optimally compatible with the text hashing network for simultaneously learning discriminative feature representation and high-quality hashing codes. The integrated text hashing network $\mathbf{Net_2}(\boldsymbol{X_2}; \boldsymbol{\theta}^{\boldsymbol{AE}}, \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{s}}; \boldsymbol{B}_{\mathbf{2}}^{\boldsymbol{s}}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{c}}; \boldsymbol{B}^{\boldsymbol{c}})$ consisting of a deep autoencoder $\boldsymbol{N}^{\boldsymbol{AE}}(\boldsymbol{X_2}; \boldsymbol{\theta}^{\boldsymbol{AE}})$, a text modality network $\boldsymbol{N}_{\mathbf{2}}^{\boldsymbol{s}}(\boldsymbol{X_2}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{s}})$ and a modality-shared network $\boldsymbol{N}_{\mathbf{2}}^{\boldsymbol{c}}(\boldsymbol{X_2}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{c}})$ is formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{\theta}^{AE}, \boldsymbol{\theta}_2^s, \boldsymbol{\theta}_2^c, \boldsymbol{B}_2^s, \boldsymbol{B}^c} \mathbf{Net_2}\left(\boldsymbol{X_2}; \boldsymbol{\theta}^{\boldsymbol{AE}}, \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{s}}; \boldsymbol{B}_{\mathbf{2}}^{\boldsymbol{s}}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{c}}; \boldsymbol{B}^{\boldsymbol{c}}\right) \\ &= \min_{\boldsymbol{\theta}^{AE}, \boldsymbol{\theta}_2^s, \boldsymbol{\theta}_2^c, \boldsymbol{B}_2^s, \boldsymbol{B}^c} \begin{pmatrix} +||\hat{\boldsymbol{N}}^{\boldsymbol{AE}}\left(\boldsymbol{X_2}; \boldsymbol{\theta}^{\boldsymbol{AE}}\right) - \boldsymbol{X_2}||^2 \\ +||\hat{\mathbf{N}}_{\mathbf{2}}^{\boldsymbol{s}}\left(\boldsymbol{X_2}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{s}}\right) - \boldsymbol{B}_{\mathbf{2}}^{\boldsymbol{s}}||^2 \\ +||\hat{\mathbf{N}}_{\mathbf{2}}^{\boldsymbol{c}}\left(\boldsymbol{X_2}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{c}}\right) - \boldsymbol{B}^{\boldsymbol{c}}||^2 \end{pmatrix} \\ &\text{s.t.} \quad \boldsymbol{B}_{\mathbf{2}}^{\boldsymbol{s}} \in \{-1,1\}^{l^s \times n}, \boldsymbol{B}^{\boldsymbol{c}} \in \{-1,1\}^{l^c \times n} \end{aligned} \tag{4}$$

where $||\hat{\boldsymbol{N}}^{\boldsymbol{AE}}(\boldsymbol{X_2}; \boldsymbol{\theta}^{\boldsymbol{AE}}) - \boldsymbol{X_2}||^2$ is the deep autoencoder loss which aims to measure the discrepancies between the network output $\hat{\boldsymbol{N}}^{\boldsymbol{AE}}(\cdot)$ of deep autoencoder with the network weights $\boldsymbol{\theta}^{\boldsymbol{AE}}$ and the input text. $||\hat{\mathbf{N}}_{\mathbf{2}}^{\boldsymbol{s}}(\boldsymbol{X_2}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{s}}) - \boldsymbol{B}_{\mathbf{2}}^{\boldsymbol{s}}||^2$ indicates the text hashing loss, which aim to extract textual features from the input texts $\boldsymbol{X_2}$ and then encode them with the text modality content codes $\boldsymbol{B}_{\mathbf{2}}^{\boldsymbol{s}}$. Similarly, $||\hat{\mathbf{N}}_{\mathbf{2}}^{\boldsymbol{c}}(\boldsymbol{X_2}; \boldsymbol{\theta}_{\mathbf{2}}^{\boldsymbol{c}}) - \boldsymbol{B}^{\boldsymbol{c}}||^2$ indicates the modality-shared hashing loss, which aim to extract textual features from the input texts $\boldsymbol{X_2}$ and then encode them with the modality-shared content code $\boldsymbol{B}^{\boldsymbol{c}}$, with $\hat{\mathbf{N}}_{\mathbf{2}}^{\boldsymbol{s}}(\cdot)$ and $\hat{\mathbf{N}}_{\mathbf{2}}^{\boldsymbol{c}}(\cdot)$ being the network outputs.

The second module of semantic index hashing aims to utilize the modality-shared content code $\boldsymbol{B}^{\boldsymbol{c}} \in \{-1,1\}^{l^c \times n}$ with clustering assignment embedding to learn a common semantic index. Upon obtaining deep features with the modality-shared network $\mathbf{N}_{\boldsymbol{i}}^{\boldsymbol{c}}(\cdot)$, the network is trained to transform the high-dimensional multimodal features into low-dimensional hashing code B^c. Then a clustering assignment matrix $\boldsymbol{H} \in \{0,1\}^{k \times n}$ is learned in cooperation with the modality-shared content code $\boldsymbol{B}^{\boldsymbol{c}}$ by minimizing the K-means error to store the common semantic information. Finally, the clustering assignment matrix $\boldsymbol{H}$ serves as the head code of the dual hashing codes to represent the shared semantic index for cross-modal retrieval. This module is formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{B}^c, \boldsymbol{H}, \boldsymbol{C}} \sum_{i=1}^{m} \left(||\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{c}}\left(\boldsymbol{X_i}; \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{c}}\right) - \boldsymbol{B}^{\boldsymbol{c}}||^2\right) + \|\boldsymbol{B}^{\boldsymbol{c}} - \boldsymbol{CH}\|^2 \\ &\text{s.t.} \quad \boldsymbol{B}^{\boldsymbol{c}} \in \{-1,1\}^{l^c \times n}, \boldsymbol{C} \in \{-1,1\}^{l^c \times k}, \\ &\qquad \boldsymbol{H} \in \{0,1\}^{k \times n}, \mathbf{1}^T \boldsymbol{H} = \mathbf{1}^T \end{aligned} \tag{5}$$

where the second term is the K-means approximation error with the clustering assignment matrix $\boldsymbol{H}$ based on the help provided by the clustering centers $\boldsymbol{C} \in \{-1,1\}^{l^c \times k}$. Here, the constant k is the number of categories. Binary constraints are imposed on $\boldsymbol{C}$ and $\boldsymbol{H}$ to ensure efficient arithmetical calculations in the Hamming space with fast XOR operations.

The third module of multimodal content hashing aims to learn the multimodal content code $\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{s}} \in \{-1,1\}^{l^s \times n}$ to maintain the cross-modal correlated content. This is achieved by introducing an integrated graph Laplacian term for each modality to preserve the modality-correlated codes among multiple modalities. One can optimize this module by minimizing the following energy function:

$$\begin{aligned} &\min_{\boldsymbol{B}_i^s, \boldsymbol{B}_j^s, \boldsymbol{L}_i, \boldsymbol{L}_j} \sum_{i=1}^{m} \begin{pmatrix} ||\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{s}}\left(\boldsymbol{X_i}; \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{s}}\right) - \boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{s}}||^2 + \\ \sum_{j>i}^{m} Tr\left(\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{s}}(\boldsymbol{L_i} + \boldsymbol{L_j})\left(\boldsymbol{B}_{\boldsymbol{j}}^{\boldsymbol{s}}\right)^T\right) \end{pmatrix} \\ &\text{s.t.} \quad \boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{s}} \in \{-1,1\}^{l^s \times n}, \boldsymbol{B}_{\boldsymbol{j}}^{\boldsymbol{s}} \in \{-1,1\}^{l^s \times n} \end{aligned} \tag{6}$$

where $\boldsymbol{L_i} \in \mathbb{R}^{n\times n}$, $\boldsymbol{L_j} \in \mathbb{R}^{n\times n}$ indicate the graph Laplacian matrices calculated on the deep multimodal features for modalities i and j, respectively, as follows:

$$\boldsymbol{L_i} = diag\left(\boldsymbol{A}_i \mathbf{1}\right) - \boldsymbol{A}_i \tag{7}$$

where $\boldsymbol{A}_i \in \mathbb{R}^{n\times n}$ represents the affinity matrix calculated on the deep features of modality i and $diag(\cdot)$ indicates a diagonal matrix with nondiagonal elements being zero.

In practice, the calculation of the affinity matrix incurs a high computational cost. To avoid this, one can adopt the anchor graph scheme, which computes the similarities between all samples and several anchor points. However, this anchor selection strategy relies heavily on the quality and number of selected anchor samples since the anchors are selected randomly. Herein, we propose a fast cross-modal affinity preservation strategy that utilizes the clustering assignment matrix $\boldsymbol{H}$ obtained by the module of semantic index hashing as guidance for the affinity matrix calculation, with the clustering assignment and affinity matrices being jointly updated during the training process. In this way, a sample only needs to calculate its similarity with the samples belonging to the same semantic category, thus achieving superior efficiency. This process can be calculated via the following function:

$$\boldsymbol{A_i^{pq}} = \begin{cases} 0, & \boldsymbol{H_p} \neq \boldsymbol{H_q} \\ \exp\left\|\bar{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{s}}\left(\boldsymbol{x_i^p}; \bar{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{s}}\right) - \bar{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{s}}\left(\boldsymbol{x_i^q}; \bar{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{s}}\right)\right\|^2 / \left(2\sigma^2\right), \\ \quad \boldsymbol{H_p} = \boldsymbol{H_q} \end{cases} \tag{8}$$

where $\bar{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{s}}(\cdot)$ and $\bar{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{s}}(\cdot)$ are the deep feature representations extracted by the last hidden fully-connected layers of the deep hashing networks for the samples of p and q. σ is the bandwidth parameter. $\boldsymbol{H_p}(\cdot)$ and $\boldsymbol{H_q}(\cdot)$ are the p-th and q-th columns in the clustering assignment matrix $\boldsymbol{H}$, respectively, which indicate the clustering labels for the samples of p and q.

By integrating (3), (4), (5), (6) of the three modules together, the overall formula is presented below:

$$\begin{aligned} &\min_{\boldsymbol{\theta^{AE}}, \boldsymbol{\theta_i^s}, \boldsymbol{\theta_i^c}, \boldsymbol{B_i^s}, \boldsymbol{B_j^s}, \boldsymbol{B^c}, \boldsymbol{H}, \boldsymbol{C}} \left\|\boldsymbol{B^c} - \boldsymbol{CH}\right\|^2 \\ &+ \sum_{i=1}^{m}\left(\mathbf{Net_i} + \sum_{j>i}^{m} Tr\left(\boldsymbol{B_i^s}\left(\boldsymbol{L_i} + \boldsymbol{L_j}\right)\left(\boldsymbol{B_j^s}\right)^T\right)\right) \\ &\text{s.t. } \boldsymbol{B_i^s} \in \{-1,1\}^{l^s\times n}, \boldsymbol{B_j^s} \in \{-1,1\}^{l^s\times n}, \\ &\quad \boldsymbol{B^c} \in \{-1,1\}^{l^c\times n}, \boldsymbol{C} \in \{-1,1\}^{l^c\times k}, \\ &\quad \boldsymbol{H} \in \{0,1\}^{k\times n}, \mathbf{1}^T\boldsymbol{H} = \mathbf{1}^T \end{aligned} \tag{9}$$

where the loss functions of integrated image hashing network $\mathbf{Net_1}$ and text hashing network $\mathbf{Net_2}$ are formulated as the (3),(4). The three modules of deep hashing network, semantic index hashing and multimodal content hashing are optimized collaboratively within the unified deep Hamming space to simultaneously bridge the semantic and heterogeneous gaps for cross-modal retrieval. Then the model could focus on the cross-modal retrieval of similar content under the guidance of the common semantic index. Finally, the semantic index $\boldsymbol{H}$ serves as the head code in the dual hashing codes of UDDH, while the integrated item $[\boldsymbol{B^c}, \boldsymbol{B_i^s}]$ of modality-shared content code $\boldsymbol{B^c}$ and multimodal content code $\boldsymbol{B_i^s}$ serve as the tail codes of UDDH. The dual deep hashing design, featuring the head code for the semantic index and tail codes for modality content, enhances cross-modal retrieval efficiency through hierarchical matching of retrieved samples.

B. Numerical Scheme to Solve UDDH

We use the popular alternating optimization scheme to obtain the numerical solution by iteratively updating each variable while fixing the others.

Optimizing the parameters $\boldsymbol{\theta^{AE}}$, $\boldsymbol{\theta_i^s}$ and $\boldsymbol{\theta_i^c}$ of the deep hashing networks: By fixing all other variables except $\boldsymbol{\theta^{AE}}$, $\boldsymbol{\theta_i^s}$ and $\boldsymbol{\theta_i^c}$, the main loss function (9) can be reduced to (3),(4).

This is a regression task that measures the discrepancies between the network outputs and the hashing codes. It is simple to optimize the parameters of the deep neural networks with standard backpropagation under the guidance of the hashing codes $\boldsymbol{B_i^s}$ and $\boldsymbol{B^c}$.

Solving the modality-shared content code $\boldsymbol{B^c}$: By fixing all other variables except $\boldsymbol{B^c}$, the main loss function (9) can be reduced to

$$\begin{aligned} &\min_{\boldsymbol{B^c}} \sum_{i=1}^{m}\left(\left\|\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{c}}\left(\boldsymbol{X_i}; \boldsymbol{\theta_i^c}\right) - \boldsymbol{B^c}\right\|^2\right) + \left\|\boldsymbol{B^c} - \boldsymbol{CH}\right\|^2 \\ &\text{s.t.} \quad \boldsymbol{B^c} \in \{-1,1\}^{l^c\times n} \end{aligned} \tag{10}$$

which can be reformulated with respect to $\boldsymbol{B^c}$ as

$$\begin{aligned} &\min_{\boldsymbol{B^c}} Tr\left((\boldsymbol{B^c})^T\boldsymbol{B^c} - 2(\boldsymbol{B^c})^T\boldsymbol{CH}\right) \\ &+ \sum_{i=1}^{m} Tr\left(-2(\boldsymbol{B^c})^T\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{c}}\left(\boldsymbol{X_i}; \boldsymbol{\theta_i^c}\right) + (\boldsymbol{B^c})^T\boldsymbol{B^c}\right) \\ &= \min_{\boldsymbol{B^c}} Tr\left((\boldsymbol{B^c})^T\left(-2\boldsymbol{CH} - 2\sum_{i=1}^{m}\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{c}}\left(\boldsymbol{X_i}; \boldsymbol{\theta_i^c}\right)\right)\right) \\ &+ Const \\ &\text{s.t.} \quad \boldsymbol{B^c} \in \{-1,1\}^{l^c\times n} \end{aligned} \tag{11}$$

where $Tr((\boldsymbol{B^c})^T\boldsymbol{B^c}) = \sum_{i=1}^{n}(\boldsymbol{b_i^c})^T\boldsymbol{b_i^c} = nl^c$, with $\boldsymbol{b_i^c}$ being the i-th column of $\boldsymbol{B^c}$. This is a constant term and can be dropped from (11), since $\boldsymbol{B^c}$ is a binary matrix. Then, one can obtain the closed-form solution of $\boldsymbol{B^c}$, which is given by

$$\boldsymbol{B^c} = \text{sign}\left(\boldsymbol{CH} + \sum_{i=1}^{m}\hat{\mathbf{N}}_{\boldsymbol{i}}^{\boldsymbol{c}}\left(\boldsymbol{X_i}; \boldsymbol{\theta_i^c}\right)\right). \tag{12}$$

Optimizing the K-means clustering centers $\boldsymbol{C}$: By fixing all other variables except $\boldsymbol{C}$, the main loss function (9) can be reduced to

$$\begin{aligned} &\min_{\boldsymbol{C}} \left\|\boldsymbol{B^c} - \boldsymbol{CH}\right\|^2 \\ &\text{s.t.} \quad \boldsymbol{C} \in \{-1,1\}^{l^c\times k} \end{aligned} \tag{13}$$

which can be rewritten as,

$$\begin{aligned}&\min_{\boldsymbol{C}} Tr\begin{pmatrix} (\boldsymbol{B^c})^T\boldsymbol{B^c} - (\boldsymbol{B^c})^T\boldsymbol{CH} \\ -\boldsymbol{H}^T\boldsymbol{C}^T\boldsymbol{B^c} + \boldsymbol{H}^T\boldsymbol{C}^T\boldsymbol{CH}\end{pmatrix}\\ &= \min_{\boldsymbol{C}} Tr\left(\boldsymbol{CHH}^T\boldsymbol{C}^T - 2\boldsymbol{CH}(\boldsymbol{B^c})^T\right) + Const\\ &\text{s.t.}\quad \boldsymbol{C}\in\{-1,1\}^{l^c\times k}.\end{aligned}\tag{14}$$

The term of $Tr(\boldsymbol{CHH}^T\boldsymbol{C}^T)$ involves discrete constraint $\boldsymbol{C}\in\{-1,1\}$, and it is difficult to directly obtain the closed-form solution of $\boldsymbol{C}$. Herein, we use the discrete cyclic coordinate descent (DCC) to update $\boldsymbol{C}$ bit by bit while satisfying the discrete constraint during optimization [48]. By decomposing the matrix $\boldsymbol{H}$ into two parts, (14) can be reformulated as:

$$\begin{aligned}&\min_{\boldsymbol{C}} Tr\left(\boldsymbol{CHH}^T\boldsymbol{C}^T - 2\boldsymbol{HQ}^T\right)\\ &= \min_{\boldsymbol{C}}\ 2\boldsymbol{h_i}\boldsymbol{H_{\neq h_i}}^T\boldsymbol{C_{\neq c_i}}^T\boldsymbol{c_i} - 2\boldsymbol{q_i}^T\boldsymbol{c_i}\\ &= \min_{\boldsymbol{C}}\ 2\left(\boldsymbol{h_i}\boldsymbol{H_{\neq h_i}}^T\boldsymbol{C_{\neq c_i}}^T - \boldsymbol{q_i}^T\right)\boldsymbol{c_i}\\ &\text{s.t. } \boldsymbol{C}\in\{-1,1\}^{l^c\times k}\end{aligned}\tag{15}$$

where $\boldsymbol{Q}^T = \boldsymbol{H}(\boldsymbol{B^c})^T$, $\boldsymbol{q_i}$ is the i-th column of $\boldsymbol{Q}$. Similarly, $\boldsymbol{h_i}$ is the i-th row of $\boldsymbol{H}$, while $\boldsymbol{H_{\neq h_i}}$ is the submatrix of $\boldsymbol{H}$ excluding $\boldsymbol{h_i}$. $\boldsymbol{c_i}$ is the i-th column of $\boldsymbol{C}$, while $\boldsymbol{C_{\neq c_i}}$ is the submatrix of $\boldsymbol{C}$, excluding $\boldsymbol{c_i}$. Finally, the optimal solution $\hat{\boldsymbol{c}}$ for one bit of $\boldsymbol{C}$ is obtained by

$$\boldsymbol{c_i} = \text{sign}\left(\boldsymbol{q_i} - \boldsymbol{C_{\neq c_i}}\boldsymbol{H_{\neq h_i}}\boldsymbol{h}_i^T\right).\tag{16}$$

Solving the common semantic index $\boldsymbol{H}$: The main loss function (9) can be reduced to the following equation with respect to $\boldsymbol{H}$ by discarding the other variables:

$$\begin{aligned}&\min_{\boldsymbol{H}} \|\boldsymbol{B^c} - \boldsymbol{CH}\|^2\\ &\text{s.t.}\quad \boldsymbol{H}\in\{0,1\}^{k\times n}, \boldsymbol{1}^T\boldsymbol{H} = \boldsymbol{1}^T.\end{aligned}\tag{17}$$

Due to the discrete constraint imposed on $\boldsymbol{H}$, it is difficult to solve this problem as a whole. Alternatively, we propose to solve each column j in $\boldsymbol{H}$ independently as follows [49]:

$$\begin{aligned}&\min_{\boldsymbol{h_j}}\ ||\boldsymbol{b^c_j} - \boldsymbol{Ch_j}||^2\\ &\text{s.t.}\quad \boldsymbol{h_j}\in\{0,1\}^{k\times 1}, ||\boldsymbol{h_j}||_1 = 1\end{aligned}\tag{18}$$

where $\boldsymbol{h_j}$ is the j-th column of $\boldsymbol{H}$. $\boldsymbol{b^c_j}$ is the j-th column of the hashing code $\boldsymbol{B^c}$. (18) is obtained by selecting a row r^* from $\boldsymbol{h_j}$, such that $\boldsymbol{h_{r^*,j}} = 1$ and $\boldsymbol{h_{r\neq r^*,j}} = 0$. Finding the optimal row r^* is equivalent to solving for the minimal distance within the Hamming space, whih is formulated as follows:

$$\begin{aligned}r^* &= \arg\min_r\ \left(Ham\left(\boldsymbol{b^c_r}, \boldsymbol{c_r}\right)\right)\\ &= \arg\min_r\ \left(l^c - \boldsymbol{b^c_r c_r}\right)\end{aligned}\tag{19}$$

where $\boldsymbol{c_r}$ is the r-th clustering center in $\boldsymbol{C}$ and l^c is the length of the modality-shared content code for the given multimodal data. $Ham(\cdot)$ denotes the distance measure in the Hamming space, which only requirs the binary bit operation and thereby efficiently reduces the computational burden. For single-label data, we can utilize (19) to learn the optimal row r^*. For multilabel data, one can learn the e labels by utilizing the top e shortest distances in the set $Ham(\boldsymbol{b^c_r}, \boldsymbol{c_r})$ sorted by distance, where e is the number of multilabel.

Algorithm 1: Unsupervised Dual Deep Hashing (UDDH) With Semantic-Index and Content-Code for Cross-Modal Retrieval.

```
Input: Multimodal data $X_i$, predefined number of
  semantic categories $k$, hashing code length $l_c$, $l^s$ and
  maximum number of iteration $Iter$.
Output: Head code of the semantic index $H$ and tail codes
  of the content codes $[B^c, B^s_i]$.
for 1 to $Iter$ do
    Fix the others and update the parameters of the deep
      hashing networks in (3), (4).
    Fix the others and update the modality-shared content
      code $B_c$ using (12).
    for 1 to $k$ do
        Fix the others and update the clustering centers $C$ for
          $K$-means using (16).
    end for
    Fix the others and update the semantic index matrix $H$
      using (19).
    Fix the others and update the multimodal content code
      $B^s_i$ using (22).
end for
Return the head code of $H$ and the tail code of
  $[B^c, B^s_i]$.
```

Solving the multimodal content code $\boldsymbol{B^s_i}$: By fixing all other variables except $\boldsymbol{B^s_i}$, the main loss function (9) can be reduced to

$$\begin{aligned}&\min_{\boldsymbol{B^s_i}}\begin{pmatrix} ||\hat{\mathbf{N}}^s_i\left(\boldsymbol{X_i};\boldsymbol{\theta^s_i}\right) - \boldsymbol{B^s_i}||^2 + \\ \sum_{j>i}^m Tr\left(\boldsymbol{B^s_i}\left(\boldsymbol{L_i}+\boldsymbol{L_j}\right)\left(\boldsymbol{B^s_j}\right)^T\right)\end{pmatrix}\\ &\text{s.t.}\quad \boldsymbol{B^s_i}\in\{-1,1\}^{l^s\times n}\end{aligned}\tag{20}$$

which can be reformulated with respect to $\boldsymbol{B^s_i}$ as

$$\begin{aligned}&\min_{\boldsymbol{B^s_i}}\begin{pmatrix} Tr\left((\boldsymbol{B^s_i})^T\boldsymbol{B^s_i} - 2(\boldsymbol{B^s_i})^T\hat{\mathbf{N}}^s_i\left(\boldsymbol{X_i};\boldsymbol{\theta^s_i}\right)\right)\\ +\sum_{j>i}^m Tr\left(\boldsymbol{B^s_j}(\boldsymbol{L_i}+\boldsymbol{L_j})^{\boldsymbol{T}}(\boldsymbol{B^s_i})^{\boldsymbol{T}}\right)\end{pmatrix}\\ &= \min_{\boldsymbol{B^s_i}} Tr\left(\left(\boldsymbol{B^s_i}\right)^T\begin{pmatrix} -2\hat{\mathbf{N}}^s_i\left(\boldsymbol{X_i};\boldsymbol{\theta^s_i}\right) + \\ \sum_{j>i}^m \boldsymbol{B^s_j}(\boldsymbol{L_i}+\boldsymbol{L_j})^T\end{pmatrix}\right) + Const\\ &\text{s.t.}\quad \boldsymbol{B^s_i}\in\{-1,1\}^{l^s\times n}\end{aligned}\tag{21}$$

where $\boldsymbol{L_i}$ and $\boldsymbol{L_j}$ are the graph Laplacian matrices. Then, one can obtain the closed-form solution of $\boldsymbol{B^s_i}$, which is given by

$$\boldsymbol{B^s_i} = \text{sign}\left(2\hat{\mathbf{N}}^s_i\left(\boldsymbol{X_i};\boldsymbol{\theta^s_i}\right) - \sum_{j>i}^m \boldsymbol{B^s_j}(\boldsymbol{L_i}+\boldsymbol{L_j})^T\right).\tag{22}$$

We summarize the implementation details of UDDH in Algorithm. 1.

C. Retrieval Complexity Analysis

The proposed UDDH exhibits superiority in the retrieval efficiency by retrieving samples from a compact search space spanned by dual hashing. Herein, we discuss theoretically the retrieval complexity of dual hashing design compared with the baselines. Because the baselines need to search for the nearest samples among all samples in the huge search space whose process is laborious, their retrieval complexity is $\mathrm{O(n_q ln)}$, where n_q, l and n are the sample size of query set, hashing code length, and sample size of retrieval set, respectively. By contrast, the retrieval complexity of UDDH is $\mathrm{O(n_q ln_a)}$, with $n_a \ll n$, where n_a is the average sample size belonging to the same semantic category. Consequently, the proposed dual deep hashing design of UDDH with head code on semantic index and tail codes on modality content makes it very efficient for cross-modal retrieval. In the retrieval stage, a query sample only needs to search for the retrieved samples with the same semantic index, thus greatly narrowing down the search space and achieving superior retrieval efficiency. This efficiency superiority of UDDH will be more and more competitive with the increasing sample size of large-scale multimodal data in the real-world applications.

IV. Experiments

We conduct experiments to evaluate the cross-modal retrieval performance of the proposed method and the compared benchmark methods. The model is evaluated on four real-world multimodal datasets and compared with traditional hashing-based and deep hashing-based cross-modal retrieval approaches. Popular metrics, including the mean average precision at top50 (MAP@50) and the Precision@10 with varying hashing code lengths, as well as the Precision-topK curve, are used to measure the performance of the tested methods.

Network Details: The proposed approach is trained by a minibatch strategy with a batch-size of 20 for both the image and text hashing networks. UDDH is trained iteratively. The hashing codes are updated once the image and text hashing networks are trained for 5 epochs, and the learning rate being decreased by a factor of 10 every three epochs. The initialized learning rates is set as 0.0003 and 0.003 for the image and text hashing networks, respectively. Furthermore, the network optimizer is Adam (Adaptive moment estimation) with 0.9 momentum and 0.0001 weight decay for both image and text hashing networks.

Baseline Methods: The popular benchmark hashing methods include joint and individual matrix factorization hashing (JIMFH) [15], robust and flexible discrete hashing (RFDH) [16], fusion similarity hashing (FSH) [17], collective matrix factorization hashing (CMFH) [19] and cross-modal discrete hashing (CMDH) [20]. Then, the cross-modal retrieval methods based on deep hashing includes unsupervised generative adversarial cross-modal hashing (UGACH) [40], deep joint-semantics reconstructing hashing (DJSRH) [41], joint-modal distribution-based similarity hashing (JDSH) [36], deep unsupervised cross-modal contrastive hashing (DUCH) [42] and deep graph-neighbor coherence preserving network (DGCPN) [43].

TABLE II
Statistics on the Tested Multimodal Datasets

Datasets	Categories	Features	Samples
Wikipedia	10	128/10	2866
Pascal-VOC	20	512/804	9963
MirFlickr	24	512/1386	25000
NUS-WIDE-10	10	500/1000	186577

A. Datasets

Four real-world multimodal datasets, including Wikipedia [1], Pascal-VOC [50], MirFlickr [51] and NUS-WIDE-10 [52], are employed for the experiments. These are multimodal datasets with image-tag or image-text pairs and largely heterogeneous feature space. Brief statistics for these datasets are shown in Table II.

The details of each dataset are shown as follows:

- *Wikipedia [1]:* This dataset is an image-text pair dataset that contains 2866 articles with their corresponding images collected from the Wikipedia website, with 10 categories. The images and texts in this dataset are represented by 128-dimensional scale invariant feature transform (SIFT) features and 10-dimensional latent Dirichlet allocation (LDA) features, respectively. In this dataset, there are 693 samples in the query subset, while 2173 samples in the retrieval subset.
- *Pascal-VOC [50]:* This dataset is an image-annotation pair dataset containing 9963 pairs of images and corresponding textual index vectors, with 20 categories. The images and texts in this dataset are represented by 512-dimensional GIST features and 804 most frequent tags. 1000 samples are randomly selected for the query subset, while the others for the retrieval subset.
- *MirFlickr [51]:* This dataset is an image-annotation pair dataset containing 25000 pairs of images and their corresponding textual index vectors, with 24 categories. The images and texts in this dataset are represented by 512-dimensional global image descriptor (GIST) features and 1386 most frequent tags. 2000 samples are randomly selected as the query subset, while the others for the retrieval subset.
- *NUS-WIDE-10 [52]:* The NUS-WIDE-10 dataset [52] contains 10 semantic concepts with 186557 images and their corresponding tag feature vectors. The images and texts are represented as 500-dimensional bag of words (BOW) features and 1000-dimensional tag vectors. 2000 samples are randomly selected as the query subset, while the others for the retrieval subset.

B. Experiment on the Effectiveness of Dual Hashing Coding for Cross-Modal Retrieval

This section tests the retrieval performance of the proposed UDDH method with that of hashing-based cross-modal retrieval methods on four real-world multimodal datasets. The experimental results in terms of the evaluation metrics of MAP@50 scores, Precision@10 scores and Precision-topK curves are reported in Tables III, IV and Fig. 2, where I → T and T →

TABLE III
THE RETRIEVAL PERFORMANCE OF OUR PROPOSED UDDH AND BENCHMARK METHODS ON THE WIKIPEDIA AND PASCAL-VOC DATASETS, EVALUATED BY MAP@50 AND PRECISION@10 ACROSS DIFFERENT HASHING CODE LENGTHS

Task	Method	Wikipedia						Pascal-VOC					
		MAP@50			Precision@10			MAP@50			Precision@10		
		32Bits	64Bits	128Bits	32Bits	64Bits	128Bits	32Bits	64Bits	128Bits	32Bits	64Bits	128Bits
I→T	CMDH	0.2141	0.2080	0.2336	0.1698	0.1650	0.1853	0.2778	0.2903	0.3011	0.2690	0.2811	0.2916
	CMFH	0.1928	0.1992	0.2065	0.1432	0.1480	0.1534	0.2675	0.2795	0.2900	0.2590	0.2707	0.2803
	FSH	0.2612	0.2590	0.2675	0.2085	0.2068	0.2136	0.2418	0.2527	0.2622	0.2342	0.2447	0.2539
	RFDH	0.2130	0.2210	0.2259	0.1591	0.1650	0.1687	0.2867	0.2996	0.3109	0.2777	0.2902	0.3010
	JIMFH	0.1937	0.1988	0.2195	0.1495	0.1534	0.1694	0.3295	0.3443	0.3572	0.3191	0.3335	0.3459
	UGACH	0.3575	0.3723	0.3716	0.3121	0.3250	0.3244	0.5237	0.5473	0.5678	0.5072	0.5301	0.5499
	DJSRH	**0.3916**	0.3884	0.4045	0.3435	0.3407	0.3548	0.5486	0.5733	0.5948	0.5313	0.5553	0.5760
	JDSH	0.3743	0.3962	0.3865	0.3255	0.3445	0.3361	0.5502	0.5750	0.5965	0.5328	0.5568	0.5777
	DUCH	0.3856	0.3963	0.4103	0.3572	0.3671	0.3801	0.5445	0.5691	0.5904	0.5274	0.5511	0.5718
	DGCPN	0.3907	0.4006	0.4206	0.3456	0.3543	0.3720	0.5553	0.5803	0.6020	0.5377	0.5620	0.5830
	UDDH	0.3908	**0.4012**	**0.4224**	**0.3610**	**0.3706**	**0.3902**	**0.5651**	**0.5906**	**0.6128**	**0.5473**	**0.5720**	**0.5934**
T→I	CMDH	0.3017	0.3149	0.3668	0.2692	0.2817	0.3281	0.2672	0.2792	0.2940	0.2487	0.2579	0.2716
	CMFH	0.2016	0.2231	0.2137	0.1436	0.1589	0.1522	0.2659	0.2758	0.2903	0.2456	0.2547	0.2682
	FSH	0.4577	0.4673	0.4926	0.4465	0.4558	0.4805	0.2573	0.2669	0.2810	0.2377	0.2466	0.2596
	RFDH	0.2064	0.2051	0.2506	0.1613	0.1602	0.1958	0.3894	0.4039	0.4253	0.3597	0.3731	0.3929
	JIMFH	0.5637	0.6279	0.6101	0.5445	0.6065	0.5893	0.3763	0.3903	0.4109	0.3476	0.3605	0.3796
	UGACH	0.5909	0.6130	0.6148	0.5763	0.5978	0.5996	0.4229	0.4386	0.4618	0.3906	0.4052	0.4266
	DJSRH	0.6135	0.6100	0.6311	0.5939	0.5935	0.6140	0.4409	0.4573	0.4815	0.4073	0.4225	0.4448
	JDSH	0.6140	0.6237	0.6267	0.5863	0.5955	0.5984	0.4512	0.4680	0.4928	0.4168	0.4323	0.4552
	DUCH	0.6040	0.6009	0.6161	0.5767	0.5738	0.5883	0.4450	0.4767	0.5019	0.4245	0.4403	0.4636
	DGCPN	0.6124	0.6301	0.6477	0.5916	0.6087	0.6257	0.4417	0.4616	0.4860	0.4111	0.4264	0.4489
	UDDH	**0.6144**	**0.6305**	**0.6541**	**0.5944**	**0.6106**	**0.6328**	**0.4694**	**0.4869**	**0.5115**	**0.4336**	**0.4498**	**0.4735**

The best scores are highlighted in bold.

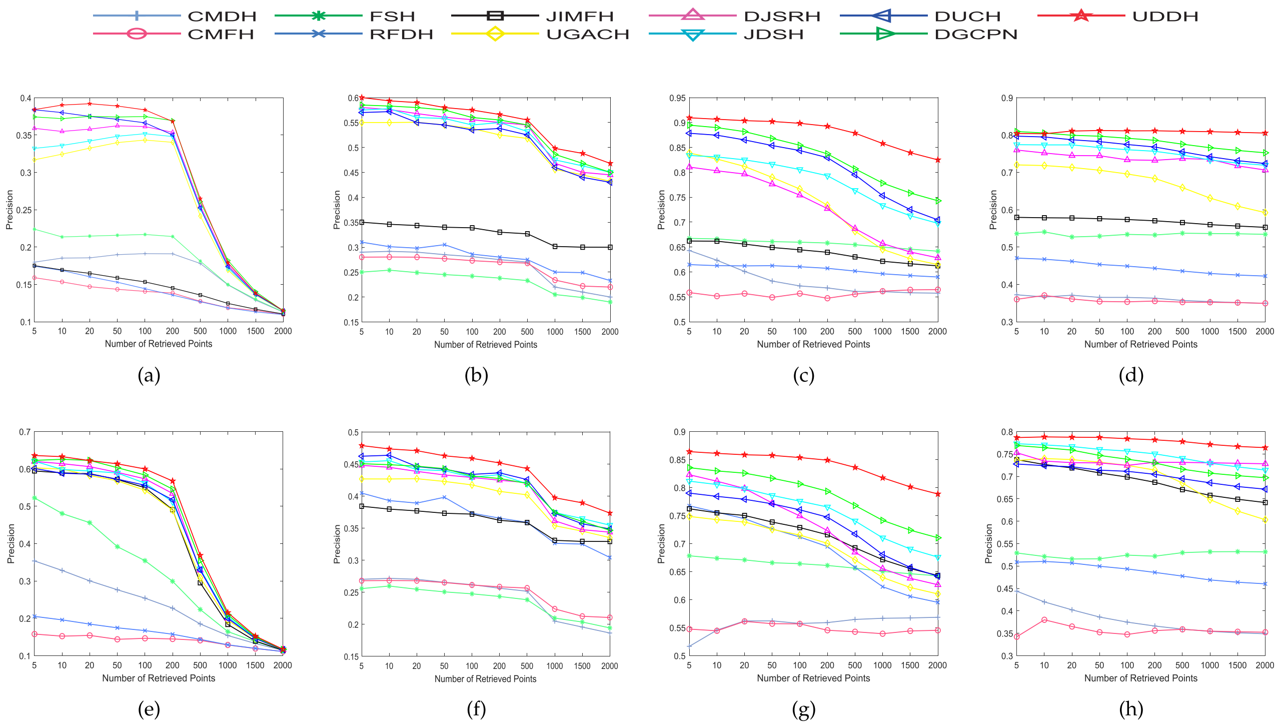
TABLE IV
THE RETRIEVAL PERFORMANCE OF OUR PROPOSED UDDH AND BENCHMARK METHODS ON THE MIRFLICKR AND NUS-WIDE-10 DATASETS, EVALUATED BY MAP@50 AND PRECISION@10 ACROSS DIFFERENT HASHING CODE LENGTHS

Task	Method	MirFlickr						NUS-WIDE-10					
		MAP@50			Precision@10			MAP@50			Precision@10		
		32Bits	64Bits	128Bits	32Bits	64Bits	128Bits	32Bits	64Bits	128Bits	32Bits	64Bits	128Bits
I→T	CMDH	0.6277	0.5966	0.6350	0.6162	0.5857	0.6234	0.4071	0.4181	0.4147	0.3596	0.3693	0.3663
	CMFH	0.5970	0.5894	0.5846	0.5631	0.5559	0.5514	0.4201	0.4146	0.4168	0.3734	0.3685	0.3705
	FSH	0.6737	0.6785	0.6873	0.6531	0.6578	0.6663	0.5485	0.5460	0.5603	0.5292	0.5268	0.5406
	RFDH	0.6429	0.6434	0.6402	0.6153	0.6158	0.6127	0.4843	0.5023	0.4988	0.4538	0.4707	0.4674
	JIMFH	0.6652	0.6694	0.6796	0.6477	0.6518	0.6617	0.5760	0.5938	0.6060	0.5497	0.5667	0.5783
	UGACH	0.8192	0.8277	0.8260	0.8199	0.8284	0.8267	0.7288	0.7396	0.7364	0.7108	0.7213	0.7182
	DJSRH	0.8153	0.8240	0.8087	0.8096	0.8182	0.8030	0.7424	0.7505	0.7687	0.7261	0.7340	0.7518
	JDSH	0.7603	0.8025	0.8375	0.7543	0.7962	0.8309	0.7527	0.7756	0.7886	0.7383	0.7607	0.7735
	DUCH	0.8120	0.8519	0.8751	0.8115	0.8514	0.8746	0.7903	0.8001	0.8021	0.7730	0.7827	0.7947
	DGCPN	**0.8702**	0.8792	0.8895	**0.8705**	0.8795	0.8898	0.7539	0.7958	0.8155	0.7445	0.7808	0.7953
	UDDH	0.8435	**0.8992**	**0.9120**	0.8385	**0.8939**	**0.9066**	**0.7913**	**0.8013**	**0.8217**	**0.7735**	**0.7833**	**0.8032**
T→I	CMDH	0.6002	0.6024	0.5891	0.5559	0.5579	0.5456	0.4335	0.4438	0.4556	0.3995	0.4090	0.4199
	CMFH	0.5813	0.5825	0.5875	0.5387	0.5398	0.5444	0.3968	0.3985	0.4035	0.3737	0.3753	0.3800
	FSH	0.6735	0.6855	0.6943	0.6536	0.6653	0.6738	0.5538	0.5566	0.5517	0.5232	0.5258	0.5212
	RFDH	0.7106	0.7373	0.7639	0.7033	0.7298	0.7561	0.5249	0.5407	0.5408	0.4952	0.5101	0.5102
	JIMFH	0.7467	0.7531	0.7683	0.7337	0.7400	0.7549	0.6952	0.7171	0.7435	0.6793	0.7007	0.7265
	UGACH	0.7188	0.7389	0.7566	0.7056	0.7253	0.7427	0.7609	0.7645	0.7543	0.7462	0.7497	0.7397
	DJSRH	0.8222	0.8177	0.8111	0.8224	0.8179	0.8113	0.7225	0.7453	0.7516	0.7057	0.7279	0.7341
	JDSH	0.7801	0.8062	0.8120	0.7738	0.7996	0.8054	0.7358	0.7519	0.7804	0.7262	0.7421	0.7702
	DUCH	0.7428	0.7687	0.7947	0.7331	0.7586	0.7843	0.7552	0.7604	0.7403	0.7386	0.7437	0.7240
	DGCPN	0.8219	0.8239	0.8372	0.8145	0.8165	0.8297	0.7565	0.7593	0.7771	0.7437	0.7465	0.7640
	UDDH	**0.8350**	**0.8583**	**0.8689**	**0.8274**	**0.8505**	**0.8610**	**0.7707**	**0.7846**	**0.8020**	**0.7572**	**0.7709**	**0.7880**

The best scores are highlighted in bold.

I indicate the image-retrieving-text and text-retrieving-image tasks, respectively. The optimal results are shown in bold for visual comparison purposes. It can be observed that the proposed UDDH method achieves uniformly superior performance over that of the other benchmark methods.

First, we evaluate the retrieval performance achieved by the proposed UDDH method and the compared benchmark methods with varying hashing code lengths on all four multimodal datasets. The experimental results are shown in Tables III and IV. It can be observed that the MAP@50 and Precision@10 values obtained by most of the methods increase as the hashing code lengths increases, due to the fact that more comprehensive information is stored by long hashing codes. In particular, the performance of UDDH is unsatisfactory when the hashing code lengths is 32-bits on the datasets of Wikipedia and MirFlickr. However, UDDH achieves uniformly superior performance and



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. Precision-topK curves produced by all cross-modal retrieval methods on multimodal datasets. The top part indicates the image-retrieving-text task conducted on the (a) Wikipedia, (b) Pascal-VOC, (c) MirFlickr and (d) NUS-WIDE-10 datasets, respectively. The bottom part indicates the text-retrieving-image task conducted on the (e) Wikipedia, (f) Pascal-VOC, (g) MirFlickr and (h) NUS-WIDE-10 datasets, respectively.

remains more stable than the other benchmark methods with hashing code lengths that are greater than or equal to 64-bits.

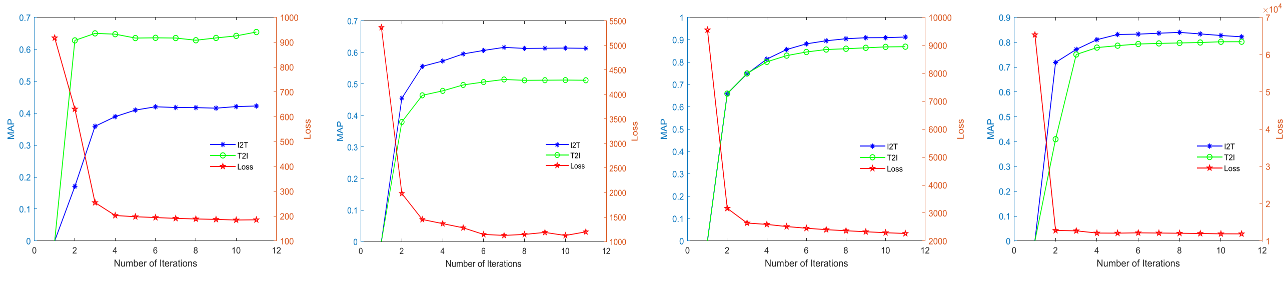
Second, it can be observed from the Tables III, IV and Fig. 2 that the performance of the traditional cross-modal retrieval methods is unsatisfactory on the four real-world multimodal datasets. The poor performance of these handcrafted-feature-based methods may result from the inaccurate similarities caused by largely heterogeneous gaps. Furthermore, their deep feature representation and binary optimization processes are completely independent and fail to learn more discriminative features and more effective hashing codes. In comparison, our proposed UDDH method utilizes deep hashing networks to extract multimodal deep features, exhibiting impressive and powerful feature representation potential, and simultaneously and jointly encode the features as binary codes. Binary optimization is optimally compatible with deep feature representation, thus yielding discriminative features and high-quality hashing codes.

Third, the cross-modal retrieval methods based on deep hashing, including UGACH [40], DJSRH [41], JDSH [36], DUCH [42] and DGCPN [43], are employed for further comparison. The experimental results are reported in Tables III, IV and Fig. 2. Our proposed UDDH also achieves superior performance than the deep hashing-based cross-modal retrieval methods. The main reason for this finding is that in the dual hashing design, the head code learning utilizes modality-shared content code with clustering assignment embedding to learn a common semantic index and modality-shared content information, and the tail code learning utilizes multimodal content codes and works with a fast cross-modal affinity preservation strategy to exploit multimodal content information. These pieces of common semantic information and modality content information are exploited and seamlessly incorporated into our deep hashing networks to simultaneously bridge the semantic and heterogeneous gaps for cross-modal retrieval. Consequently, the dual hashing design of UDDH can also achieve superior retrieval performance. Furthermore, the learning of deep feature extraction, binary optimization, the common semantic index and the modality content codes are integrated into a unified model with collaborative optimization so that they can benefit from each other and further improve the final retrieval performance.

Finally, we present the convergence analysis of the proposed UDDH. The numerical scheme of UDDH utilizes the standard alternating optimization scheme to obtain numerical solutions by iteratively updating each variable while fixing the others. As we can see, the designed numerical scheme converges very quickly and the MAP@50 scores become stable within just a few iterations, as shown in Fig. 3. Therefore, the optimization scheme of our proposed UDDH achieves superior efficiency. In our experiments, we use the solution at iteration 10 as our final experimental results.

C. Experiment on the Effectiveness Test of Samplewise Retrieval of Similar Content

Our approach is different from the traditional supervised or unsupervised cross-modal retrieval methods, which transform multimodal features into a common semantic space or latent



(a) (b) (c) (d)

Fig. 3. The values obtained for the loss function and the MAP@50 metric as the number of iterations on the (a) Wikipedia, (b) Pascal-VOC, (c) MirFlickr and (d) NUS-WIDE-10 datasets, respectively. The left ordinate is the MAP@50 value, while the right ordinate is the loss function value. The proposed UDDH converges very quickly and the solution becomes stable within just a few iterations.

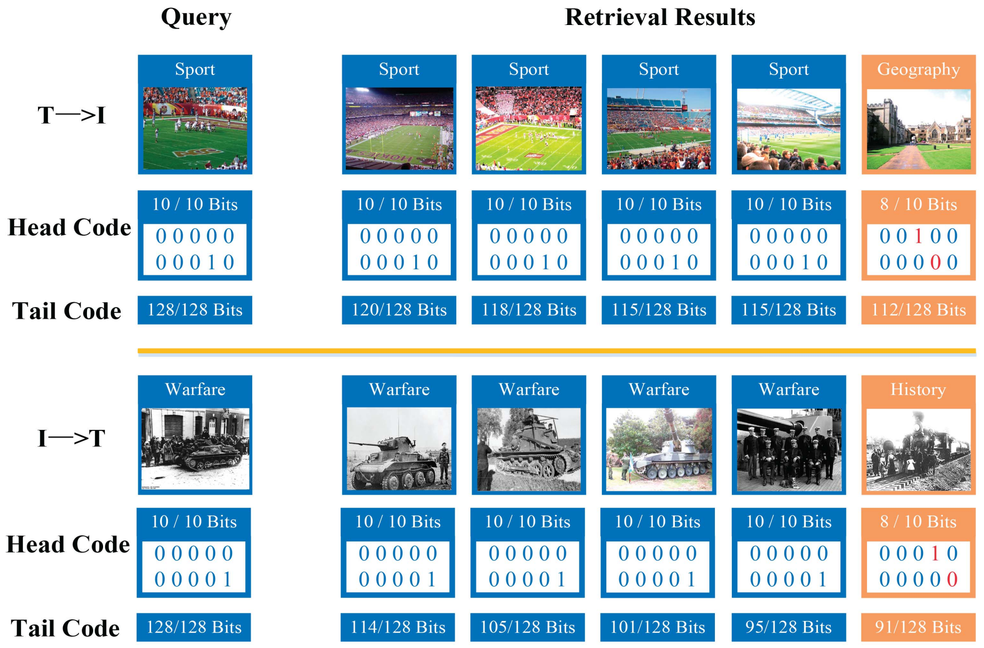


Fig. 4. Illustrative examples on the top five retrieval results obtained by the UDDH on the Wikipedia dataset. One can verify that the top retrieved samples well match the query samples with very similar image content.

space. In this space, the multimodal features are encoded as identical hashing codes in general. For example, traditional supervised methods mainly focus on semanticwise cross-modal retrieval and easily lose sample-specific and modality-specific properties. Herein, our proposed UDDH method focuses on samplewise similar content cross-modal retrieval with semantic guidance by preserving samplewise multimodal similarities. Therefore, we illustrate the top five retrieval results with their associated semantic labels in Fig. 4. For visual convenience, we replace the Wikipedia text samples with their corresponding images. The successful and failed retrieval results are highlighted in blue and orange, respectively. The number in the figure indicates the number of matched hashing codes out of all hashing codes. We can see that the top retrieved samples match the query samples well based on the image content. Although several retrieved semantic labels are incorrect, their content is similar to the query content. Consequently, after obtaining the head code of the semantic index, the proposed UDDH method can exploit the samplewise correlated content under the guidance of the same semantic index, then focusing on the samplewise retrieval of similar content.

V. Conclusion

The existing unsupervised hashing-based cross-modal retrieval methods rely on accurate sample similarity. Such approach requires searching for the nearest samples within a vast search space, resulting in intensive computation. In this paper, we have proposed UDDH which encodes each input multimodal sample by dual binary codes with a head code for the semantic index and tail codes for the modality content. This dual hashing design makes the proposed approach very efficient and effective for cross-modal retrieval. UDDH utilizes deep hashing networks to extract multimodal features and jointly encode them as binary codes, thus yielding discriminative features and high-quality hashing codes. Experimental results have demonstrated that the proposed UDDH method achieves superior performance in cross-modal retrieval tasks, by exploiting common semantic information and modality content information to simultaneously bridge the semantic and heterogeneous gaps.

References

[1] N. Rasiwasia et al., “A new approach to cross-modal multimedia retrieval,” in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[2] Y. Wei et al., “Modality-dependent cross-media retrieval,” *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 1–13, 2016.

[3] F. Shang, H. Zhang, J. Sun, and L. Liu, “Semantic consistency cross-modal dictionary learning with rank constraint,” *J. Vis. Commun. Image Representation*, vol. 62, pp. 259–266, 2019.

[4] X. Lu, L. Zhu, Z. Cheng, J. Li, X. Nie, and H. Zhang, “Flexible online multi-modal hashing for large-scale multimedia retrieval,” in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1129–1137.

[5] H. Xiong et al., “A generalized method for binary optimization: Convergence analysis and applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4524–4543, Sep. 2022.

[6] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Image-text embedding learning via visual and textual semantic reasoning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 641–656, Jan. 2023.

[7] J. Wei, Y. Yang, X. Xu, X. Zhu, and H. T. Shen, “Universal weighting metric learning for cross-modal retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6534–6545, Oct. 2022.

[8] X. Xu, K. Lin, Y. Yang, A. Hanjalic, and H. T. Shen, “Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3030–3047, Jun. 2022.

[9] J. Marin et al., “Recipe1m : A dataset for learning cross-modal embeddings for cooking recipes and food images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 187–203, Jan. 2021.

[10] Y. Feng, S. Ji, Y.-S. Liu, S. Du, Q. Dai, and Y. Gao, “Hypergraph-based multi-modal representation for open-set 3D object retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2206–2223, Apr. 2024, doi: 10.1109/TPAMI.2023.3332768.

[11] A. J. Wang, P. Zhou, M. Z. Shou, and S. Yan, “Enhancing visual grounding in vision-language pre-training with position-guided text prompts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3406–3421, May 2024, doi: 10.1109/TPAMI.2023.3343736.

[12] G. Song, S. Wang, Q. Huang, and Q. Tian, “Harmonized multimodal learning with Gaussian process latent variable models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 858–872, Mar. 2021.

[13] Z. Zhang et al., “Universal multimodal representation for language understanding,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9169–9185, Jul. 2023.

[14] X. Dong et al., “Entity-graph enhanced cross-modal pretraining for instance-level product retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13117–13133, Nov. 2023.

[15] D. Wang, Q. Wang, L. He, X. Gao, and Y. Tian, “Joint and individual matrix factorization hashing for large-scale cross-modal retrieval,” *Pattern Recognit.*, vol. 107, pp. 1–12, 2020.

[16] D. Wang, Q. Wang, and X. Gao, “Robust and flexible discrete hashing for cross-modal similarity search,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2703–2715, Oct. 2018.

[17] H. Liu, R. Ji, F. Huang, and B. Zhang, “Cross-modality binary code learning via fusion similarity hashing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7380–7388.

[18] J. Song, Y. Yang, Y. Yang, Z. Huang, and S. H. Tao, “Inter-media hashing for large-scale retrieval from heterogeneous data sources,” in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 785–796.

[19] G. Ding, Y. Guo, and J. Zhou, “Collective matrix factorization hashing for multimodal data,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2075–2082.

[20] V. E. liong, J. Lu, and Y.-P. Tan, “Cross-modal discrete hashing,” *Pattern Recognit.*, vol. 79, pp. 114–129, 2018.

[21] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, “Linear cross-modal hashing for efficient multimedia search,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 143–152.

[22] J. Zhou, G. Ding, and Y. Guo, “Latent semantic sparse hashing for cross-modal similarity search,” in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 415–424.

[23] M. Hu, Y. Yang, F. Shen, N. Xie, R. Hong, and H. T. Shen, “Collective reconstructive embeddings for cross-modal hashing,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2770–2784, Jun. 2019.

[24] D. Wang, X. Gao, X. Wang, and L. He, “Label consistent matrix factorization hashing for large-scale cross-modal similarity search,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2466–2479, Oct. 2019.

[25] M. Lin et al., “Fast class-wise updating for online hashing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2453–2467, May 2022.

[26] X. Lu, L. Zhu, J. Li, H. Zhang, and H. T. Shen, “Efficient supervised discrete multi-view hashing for large-scale multimedia search,” *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2048–2060, Aug. 2020.

[27] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, “Fast supervised discrete hashing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 490–496, Feb. 2018.

[28] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, “MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.

[29] W. Chen, W. Wang, L. Liu, and M. S. Lew, “New ideas and trends in deep multimodal content understanding: A review,” *Neurocomputing*, vol. 426, pp. 195–215, 2021.

[30] Q.-Y. Jiang and W.-J. Li, “Deep cross-modal hashing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3232–3240.

[31] Y.-W. Zhan, X. Luo, Y. Wang, and X.-S. Xu, “Supervised hierarchical deep hashing for cross-modal retrieval,” in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 3386–3394.

[32] L. Jin, Z. Li, and J. Tang, “Deep semantic multimodal hashing network for scalable image-text and video-text retrievals,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1838–1851, Apr. 2023.

[33] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, “Deep visual-semantic hashing for cross-modal retrieval,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1445–1454.

[34] L. Jin, K. Li, Z. Li, F. Xiao, G.-J. Qi, and J. Tang, “Deep semantic-preserving ordinal hashing for cross-modal similarity search,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1429–1440, May 2019.

[35] C. Li, C. Deng, L. Wang, D. Xie, and X. Liu, “Coupled CycleGAN: Unsupervised hashing network for cross-modal retrieval,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 176–183.

[36] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, “Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval,” in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1379–1388.

[37] M. Li and H. Wang, “Unsupervised deep cross-modal hashing by knowledge distillation for large-scale cross-modal retrieval,” in *Proc. Int. Conf. Multimedia Retrieval*, 2021, pp. 183–191.

[38] D. Hu, F. Nie, and X. Li, “Deep binary reconstruction for cross-modal hashing,” *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 973–985, Apr. 2019.

[39] J. Huang, C. Min, and L. Jing, “Unsupervised deep fusion cross-modal hashing,” in *Proc. Int. Conf. Multimodal Interaction*, 2019, pp. 358–366.

[40] J. Zhang, Y. Peng, and M. Yuan, “Unsupervised generative adversarial cross-modal hashing,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 539–546.

[41] S. Su, Z. Zhong, and C. Zhang, “Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3027–3035.

[42] G. Mikriukov, M. Ravanbakhsh, and B. Demir, “Unsupervised contrastive hashing for cross-modal retrieval in remote sensing,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 1–8.

[43] J. Yu, H. Zhou, Y. Zhan, and D. Tao, “Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 4626–4634.

[44] D. Wang, C. Zhang, Q. Wang, Y. Tian, L. He, and L. Zhao, “Hierarchical semantic structure preserving hashing for cross-modal retrieval,” *IEEE Trans. Multimedia*, vol. 25, pp. 1217–1229, 2023.

[45] X. Shen, H. Zhang, L. Li, and L. Liu, “Attention-guided semantic hashing for unsupervised cross-modal retrieval,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[48] F. Shen, C. Shen, W. Liu, and H. Tao Shen, “Supervised discrete hashing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 37–45.
[49] L. Huang, H.-Y. Chao, and C.-D. Wang, “Multi-view intact space clustering,” *Pattern Recognit.*, vol. 86, pp. 344–353, 2019.
[50] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” in *Proc. Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[51] M. J. Huiskes and M. S. Lew, “The MIR flickr retrieval evaluation,” in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.
[52] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “NUS-WIDE: A real-world web image database from National University of Singapore,” in *Proc. ACM Conf. Image Video Retrieval*, 2009, pp. 1–9.

Bin Zhang received the BS and MS degrees from the School of Information Science and Engineering, Shandong Normal University, Jinan, Shandong, China, in 2011 and 2015, respectively, and the PhD degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China. He is a postdoctoral researcher with the Guangdong Institute of Intelligence Science and Technology. His current research interests include large-scale clustering, feature matching, cross-modal retrieval, medical image processing, and bioinformatics.

Yue Zhang received the PhD degree in computer science from Hong Kong Baptist University, Hong Kong SAR, China, in 2017. She is an associate professor with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. Her research interests include bioinformatics and Big Data mining.

Junyu Li received the BS and MS degrees from the Guangdong University of Technology, Guangzhou, China, in 2017 and 2020, respectively. He is currently working toward the the PhD degree in computer science and engineering from the South China University of Technology, Guangzhou. His research interests include machine learning and image processing.

Jiazhou Chen (Member, IEEE) received the BS degree in computer science and technology from Jiaying University, Meizhou, China, the MS degree in software engineering from the Guangdong University of Technology, Guangzhou, China, and the PhD degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou. His current research interests include bioinformatics, machine learning, computer vision, image processing.

Tatsuya Akutsu (Senior Member, IEEE) received the BE and ME degrees in aeronautics and the DE degree in information engineering from The University of Tokyo, Tokyo, Japan, 1984, 1986, and 1989, respectively. He has been a professor with the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan, since 2001. His research interests include bioinformatics and discrete algorithms.

Yiu-Ming Cheung (Fellow, IEEE) received the PhD degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong SAR, China. He is currently a chair professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, and visual computing. He is the founding chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He is the editor-in-chief of *IEEE Transactions on Emerging Topics in Computational Intelligence*. Also, he serves as an associate editor for *IEEE Transactions on Cybernetics*, *IEEE Transactions on Cognitive and Developmental Systems*, *IEEE Transactions on Neural Networks and Learning Systems* (2014-2020), *Pattern Recognition*, *Knowledge and Information Systems*, and *Neurocomputing*, as well as the guest editor in several international journals. He is a fellow of the AAAS, IET, and BCS.

Hongmin Cai (Senior Member, IEEE) received the BS and MS degrees in mathematics from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the PhD degree in applied mathematics from Hong Kong University, in 2007. He is a professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. From 2005 to 2006, he was a research assistant with the Center of Bioinformatics, Harvard University, and Section for Biomedical Image Analysis, University of Pennsylvania. His areas of research interests include biomedical image processing and omics data integration.