

Lightweight Contrastive Distilled Hashing for Online Cross-modal Retrieval

Jiaxing Li¹, Lin Jiang², Zeqi Ma³, Kaihang Jiang³, Xiaozhao Fang^{4,*}, Jie Wen⁵

¹School of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China

²School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

³School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong SAR, China

⁴School of Automation, Guangdong University of Technology, Guangzhou 510006, China

⁵Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518067, China

jiaxing.li.cs@gmail.com, linn_jiang@163.com, 22123638r@connect.polyu.hk, jkh1650290810@163.com, xzhfang168@126.com, jiewen_pr@126.com

Abstract

Deep online cross-modal hashing has gained much attention from researchers recently, as its promising applications with low storage requirement, fast retrieval efficiency and cross modality adaptive, etc. However, there still exists some technical hurdles that hinder its applications, e.g., 1) how to extract the coexistent semantic relevance of cross-modal data, 2) how to achieve competitive performance when handling the real time data streams, 3) how to transfer the knowledge learned from offline to online training in a lightweight manner. To address these problems, this paper proposes a lightweight contrastive distilled hashing (LCDH) for cross-modal retrieval, by innovatively bridging the offline and online cross-modal hashing by similarity matrix approximation in a knowledge distillation framework. Specifically, in the teacher network, LCDH first extracts the cross-modal features by the contrastive language-image pre-training (CLIP), which are further fed into an attention module for representation enhancement after feature fusion. Then, the output of the attention module is fed into a FC layer to obtain hash codes for aligning the sizes of similarity matrices for online and offline training. In the student network, LCDH extracts the visual and textual features by lightweight models, and then the features are fed into a FC layer to generate binary codes. Finally, by approximating the similarity matrices, the performance of online hashing in the lightweight student network can be enhanced by the supervision of coexistent semantic relevance that is distilled from the teacher network. Experimental results on three widely used datasets demonstrate that LCDH outperforms some state-of-the-art methods.

Introduction

In the past decades, the volume of multimedia data has explosively grown in an exponential way (Jiang et al. 2024b; Cui et al. 2023b; Jiang et al. 2023b; Xu et al. 2023; Ling et al. 2023; Cui et al. 2023a; Chen et al. 2024), especially in some artificial intelligence (AI)-based applications. Under this circumstance, it is much more difficult to provide the real time retrieval services for the cross-modal data with such huge volume. As a powerful model in AI-based applications, large language model (LLM) is benefited from

the increasing volume of multimedia data, and has shown its great potentials in representation learning, neural language processing and generation, etc. Training of LLMs involves a large number of parameters, and thus they can be applied to many complex scenarios for real time prediction or processing (Zhu et al. 2024). However, the devices that handling the cross-modal retrieval tasks in AI applications may not be powerful in performance, especially for the mobile devices. Thus, a lightweight retrieval framework supervised by the large training model for multimedia data is with urgent need.

However, the application of lightweight online cross-modal hashing that is supervised by large-scale model training in the knowledge distillation framework, is hindered by some technical hurdles. Firstly, it is difficult in cross-modal retrieval to extract coexistent semantic relevance of data from different modalities, as there exists huge semantically heterogeneous gaps between different modalities. Secondly, the feature representation of coexistent semantic relevance on similarity is essential to effectively achieve better performances in hashing-based cross-modal retrieval, but most of the existing methods may fail to adequately obtain the effective representation. Finally, bridging the connection between the offline hashing in teacher network and online hashing in the lightweight student network under the knowledge distillation framework is also challenging.

To overcome the above-mentioned technical hurdles, this paper proposes a lightweight contrastive distilled hashing (LCDH) for online cross-modal retrieval. LCDH follows the research trend that combining traditional methods and modern neural networks to provide a lightweight retrieval framework for AI applications (Zhu et al. 2024). Specifically, LCDH first fuses the cross-modal features extracted by CLIP, and the fused feature is further fed into an attention module to enhance feature representation in the teacher network. Then, for aligning the sizes of similarity matrices in offline and online hashing, the output of the attention module is fed into a FC layer. In the student network, LCDH extracts image and text features by lightweight models, and then the features are fed into a FC layer to obtain binary codes, respectively. Finally, the performance of online cross-modal hashing in the lightweight student network can be boosted under the supervision of the coexistent semantic relevance which is learned from the teacher network, after similarity

*Corresponding Author: Xiaozhao Fang.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

approximation by bridging the online and offline similarity matrices. Main contributions are summarized as follows.

- Different from the existing cross-modal hashing methods, the proposed LCDH can effectively extract the co-existent semantic relevance of cross-modal data for the offline cross-modal hashing in teacher network, and then the co-existent semantic relevance is distilled and transferred for guiding the online cross-modal retrieval in the lightweight student network.
- To make the lightweight applications of LLMs can be efficiently deployed into the client devices, LCDH first adopts CLIP and attention modules to effectively preserve the co-existent semantic relevance on modalities. After the feature extraction and enhancement in the teacher network, the features are fused to obtain discriminative hash codes that can further supervise the online training in lightweight student network by bridging with the offline and online similarity matrices.
- Extensive experimental results on three widely used datasets demonstrate that LCDH can deliver the best and stable performance consistently, compared with other state-of-the-art baseline methods.

Related Works

Offline Cross-modal Hashing

Offline cross-modal hashing methods attempt to train hash codes and functions with full database in a batch-based manner, which suffers from the limitation that cannot handle the real time data chunks (Huang et al. 2023). For example, deep cross-modal hashing (DCMH) adopts the end-to-end strategy to jointly learn hash codes and functions for cross-modal data (Jiang and Li 2017). Self-supervised adversarial hashing (SSAH) leverages two adversarial-based models to maximize correlations on semantic between modalities (Li et al. 2018). Unsupervised knowledge distillation (UKD) hashing constructs similarity matrix in the unsupervised teacher network for guiding the hash codes and functions learning in the supervised student network (Hu et al. 2020). Deep cross-modal hashing with hash functions and unified hash code jointly learning (DCUCH) generates hash codes and learns hash functions asymmetrically (Tu et al. 2022). Deep adaptive enhanced hashing (DAEH) uses adaptive optimization method and discriminative similarity measurement to learn hash functions (Shi et al. 2022). Semi supervised knowledge distillation for cross-modal hashing (SKDCH) learns similarity as knowledge in a semi-supervised method in teacher network to guide the supervised cross-modal hashing in student network (Su et al. 2023). Coding self-representative and label-relaxed hashing (CSLRH) enhance the class-specific feature representations and the semantic discriminability, while reduce the noise interference (Jiang et al. 2024a). CLIP based knowledge distillation hashing (CKDH) leverages a knowledge distillation framework to transfer the deep semantic similarity learned by CLIP to the lightweight cross-modal hashing in the student network (Li et al. 2024).

Online Cross-modal Hashing

Unlike offline cross-modal hashing methods, online cross-modal hashing methods train hash codes and functions in a real-time updating manner, where hash functions are updated by hash codes trained by the continually coming data chunks. For example, hadamard code book based online hashing (HCOH) leverages hadamard code book to generate hadamard matrix for online hash codes and functions learning (Lin et al. 2018). Online latent semantic hashing (OLSH) maps discrete labels to a continuous latent semantic space which is bridged with binary codes for learning hash codes (Yao et al. 2019). Online collective matrix factorization hashing (OCMFH) uses the collective matrix factorization method to decompose new data chunk’s features and existing features into a common latent space that is generated by the features and hash function simultaneously (Wang et al. 2020). Label embedding online hashing (LEMON) leverages similarity factorization hashing method for addressing the online hashing retrieval problem, by reconstructing the similarity matrices using label information for newly coming online data chunks (Wang, Luo, and Xu 2020). Supervised discrete online hashing (SDOH) embeds semantic label into the common latent space for parallel calculating the common representation of the newly coming data from multiple modalities, and then generates hash codes by the continuous common latent representation (Liu et al. 2022). Random online hashing (ROH) proposes a linear bridging strategy to simplify the similarity factorization problem into a linear optimization problem, and then proposes a MED embedding methods to learn features and preserve significant semantic information into hash codes (Jiang et al. 2023a).

The Proposed Method

Problem Statement

Assuming that there have n pairs of image-text instances in the training datasets $\mathcal{O} = \{o_i\}_{i=1}^n$. Let v_j be the j -th image instance and b_j be the j -th text instance, each pair of instances in \mathcal{O} can be represented as $o_j = \{v_j, b_j\}$. The label matrix for \mathcal{O} can be denoted as $L \in \{0, 1\}^{n \times c}$, and c is the number of categories. The deep semantic features that are extracted from the image feature encoder in the offline manner can be denoted as $F_I \in \mathbb{R}^{n \times d_I}$, while the deep semantic features that are extracted from the text feature encoder in the offline manner can be denoted as $F_T \in \mathbb{R}^{n \times d_T}$. Similarly, the deep semantic features that are extracted from the image and text feature encoders at round t in the online manner can be denoted as $F_I^{(t)} \in \mathbb{R}^{n \times d_I}$ and $F_T^{(t)} \in \mathbb{R}^{n \times d_T}$, respectively. In addition, d_I is the dimension of extracted image features, and d_T is the dimension of extracted text features. $\|\cdot\|_F^2$ is the Frobenius norm, in which $\|M\|_F^2 = \text{Tr}(M^T M)$ and $\text{Tr}(M)$ is the trace of matrix M .

Moreover, training set \mathcal{O} is divided into two parts, in which former n_o pairs instances are utilized for offline training in the teacher network and the rest n_t pairs instances are utilized for online training in the student network. That is, $v_* = \{v_*^o, v_*^t\}$, $b_* = \{b_*^o, b_*^t\}$, and $n = n_o + n_t$. Similarly, the label matrices for instances in offline cross-modal

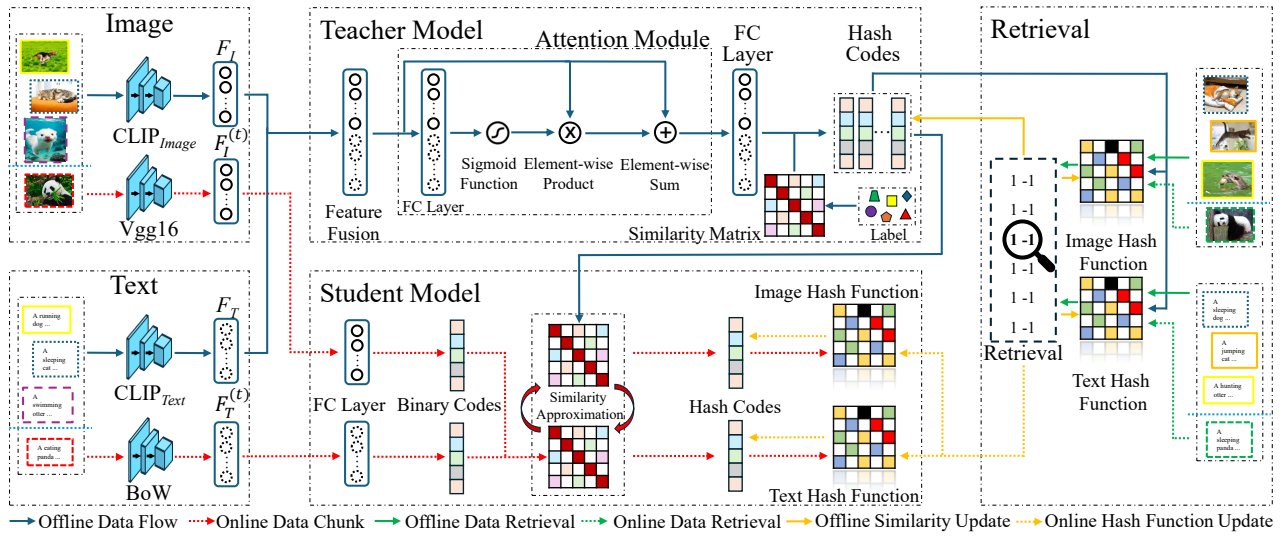


Figure 1: The framework of the lightweight contrastive distilled hashing (LCDH) for online cross-modal retrieval.

hashing is $L \in \{0, 1\}^{n_o \times c}$, and in online cross-modal hashing is $L^{(t)} \in \{0, 1\}^{n_t \times c}$. After deep feature extraction, the deep semantic features for image and text instances in offline cross-modal hashing are fused and fed into an attention module for further cross-modal semantic preservation. The output of the attention module is fed into a fully-connected (FC) layer, and then the hash codes $B' \in \{-1, +1\}^{n_t \times r}$ is generated for constructing the similarity matrix. Notably, r is the length of hash codes. The hash codes for image and text instances in online cross-modal hashing can be represented as $V^{(t)} = \{v_{*,i}^t\}_{i=1}^{n_t} \in \{-1, +1\}^{n_t \times r}$ and $B^{(t)} = \{b_{*,i}^t\}_{i=1}^{n_t} \in \{-1, +1\}^{n_t \times r}$, where n_t is the size of new coming data chunk.

Feature Extraction

The CLIP model takes cross-modal data as input to extract deep semantic features for offline cross-modal hashing in teacher network, while the Vgg16 and BoW models are utilized for extracting deep semantic features for online cross-modal hashing in lightweight student network. Notably, to make the semantic similarity preserved in the teacher network can supervise the online training in student network, the length of output from the FC layer is set to n_t . As such, the size of similarity matrix S generated by hash codes learned in the teacher network and similarity matrix $S^{(t)}$ updated in the student network is same (i.e., $n_t \times n_t$).

Offline feature extraction: In the teacher network, the CLIP model takes $\{o_i\}_{i=1}^{n_o}$ as input, and outputs $F_I = \{f_i^1, f_i^2, \dots, f_i^{n_o}\} \in \mathbb{R}^{n_o \times d_I}$ and $F_T = \{f_t^1, f_t^2, \dots, f_t^{n_o}\} \in \mathbb{R}^{n_o \times d_T}$, where d_I and d_T are the dimensions of F_I and F_T , respectively. Specifically, $f_i^j = CLIP_{Image}(v_j)$ and $f_t^j = CLIP_{Text}(b_j)$, where $j \in [1, n_o]$. Then, the fusion of features F_I and F_T is input into the attention module as F_f . That is, $F_f = \{f_i^1 f_t^1, f_i^2 f_t^2, \dots, f_i^{n_o} f_t^{n_o}\}$. The deep features F_f is then fed into a FC layer to generate binary code $B \in$

$\{-1, +1\}^{n_o \times r}$. The filter process in attention module can be defined as follows:

$$M_f = Sigmoid(WF_f + D), \quad (1)$$

where $M_f \in \mathbb{R}^{n_o \times d}$ is the mask weight matrix and d is the dimension of the fused feature. Moreover, D and W are the parameters for the FC layer, and $Sigmoid$ is the activation function. The output of attention module is

$$A_f = F_f + M_f \odot F_f, \quad (2)$$

where \odot indicates element wise product and $A_f \in \mathbb{R}^{n_o \times d}$. Finally, A_f is fed into the FC layer for generating hash codes (i.e., $B' \in \{-1, +1\}^{n_t \times r}$) which are well preserved the modality-specific similarity. As such, B' can be further utilized for constructing the similarity matrix S to supervise the training in online cross-modal hashing of student network.

Online Feature Extraction: In the student network, we adopt lightweight feature extraction models for online training. Specifically, we utilize Vgg16 model to extract deep semantic feature $F_I^{(t)}$ for image instances, and we utilize BoW model to extract deep semantic feature $F_T^{(t)}$ for text instances. That is, $F_I^{(t)} = Enc_{Vgg16}(\{v_j\}_{j=1}^{n_t})$ and $F_T^{(t)} = Enc_{BoW}(\{b_j\}_{j=1}^{n_t})$, where $Enc_{*}(\cdot)$ is the encoder. After feature extracting, $F_I^{(t)}$ and $F_T^{(t)}$ are fed into the FC layer for further generating binary codes.

Similarity Preserving

Similarity preservation in teacher network: In the teacher network, the output of FC layer is further supervised by the label information in similarity matrix. Following (Zhang and Li 2014), we randomly select n_t instances from offline training set $\{v_*, b_*\}$ to initialize the similarity matrix as follows:

$$S = 2GG^T - \mathbf{1}\mathbf{1}^T, \quad (3)$$

where L' is the label matrix of the selected n_t instances, $G \in \{0, 1\}^{n_t \times c}$ is generated by label matrix $L' \in \{0, 1\}^{n_t \times c}$ and

$G = L'/\|L'\|_F^2$. Then, the S in the offline cross-modal hashing of teacher network can be constructed as follows.

$$\mathcal{L}_T = \min_{B'} \|rS - B'^T B'\|_F^2, \quad (4)$$

Similarity preservation in student network: Normally, in the online cross-modal hashing of student network, the semantic similarity between cross-modal data can be embedded into binary codes as follows (Liu et al. 2012):

$$\min_{V^{(t)}, B^{(t)}} \|rS^{(t)} - V^{(t)} B^{(t)T}\|_F^2, \quad (5)$$

where $S^{(t)} \in \{-1, +1\}^{n_t \times n_t}$ is the semantic similarity matrix. For the i -th and j -th online instances, if they share at least 1 common label $S_{i,j}^{(t)} = 1$, otherwise $S_{i,j}^{(t)} = 0$. However, this method may fail to preserve the fine-grained semantic similarity for multiple label data. At the same time, this method may also lead to large overheads on storage and computation during training, which goes against the original intention of hashing-based cross-modal retrieval. To overcome the above issues and make $S^{(t)}$ available for online scenarios, this paper adopts logistic regression to preserve the similarity for cross-modal data. Let $g(\Omega_{i,j}^{(t)}) = G_{i,j}^{(t)} = \frac{1}{1+e^{-\Omega_{i,j}^{(t)}}}$, where $\Omega_{i,j}^{(t)} = \frac{\omega}{r} B_{i,*}^{(t)} V_{j,*}^{(t)}$ and ω is a scaling factor of tuning. Then, we can define the likelihood probability function for $S^{(t)}$ as follows:

$$p(S_{i,j}^{(t)}|B^{(t)}) = \begin{cases} G_{i,j}^{(t)}, & \text{if } S_{i,j}^{(t)} = 1, \\ 1 - G_{i,j}^{(t)}, & \text{otherwise.} \end{cases} \quad (6)$$

where the similarity matrix $S^{(t)}$ is constructed by label matrix (i.e., $S^{(t)} = L^{(t)} L^{(t)T}$ and $L^{(t)} \in \{0, 1\}^{n_t \times c}$) and the value of elements in $S^{(t)}$ is binary as follows:

$$S_{i,j}^{(t)} = \begin{cases} 1, & \text{if the } i\text{-th and } j\text{-th samples are similar,} \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

To obtain the maximum likelihood estimation, a log-likelihood similarity preserving term is adopted as follows:

$$\log L = \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \log p(S_{i,j}^{(t)}|B^{(t)}, V^{(t)}). \quad (8)$$

Finally, the maximum log-likelihood similarity preserving problem can be defined as follows:

$$\max_{S^{(t)}} \mathcal{L}_{log} = \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} [S_{i,j}^{(t)} \Omega_{i,j}^{(t)} - \log(1 + e^{\Omega_{i,j}^{(t)}})]. \quad (9)$$

Knowledge Distillation

After offline cross-modal hashing in teacher network, the semantic similarity is first initialized by Eq. (3) and then constructed by Eq. (4). As such, the deep semantic cross-modal similarity extracted by CLIP and attention module in teacher network can be well preserved by Eq. (9), and further transferred to the online training in lightweight student network by the knowledge distillation framework in LCDH. In

the lightweight student network, the image and text features are extracted by Vgg16 and BoW models from the online training set in a lightweight manner, respectively. Then, the hash codes and functions are updated by the maximum log-likelihood similarity preservation. Meanwhile, the similarity matrix $S^{(t)}$ is updated by S as follows:

$$\min_{S^{(t)}} \mathcal{L}_{KD} = \|S - S^{(t)}\|_F^2. \quad (10)$$

As such, the deep semantic similarity extracted from CLIP and attention module is distilled and transferred to $S^{(t)}$, which guides the online cross-modal hashing in the lightweight student network.

Objective Function

By integrating Eq. (4), (9) and (10), the objective function of LCDH can be formulated as follows:

$$\min_{B', V^{(t)}, B^{(t)}, S^{(t)}} \mathcal{L} = -\mathcal{L}_{log} + \lambda_1 \mathcal{L}_T + \lambda_2 \mathcal{L}_{KD}, \quad (11)$$

where λ_1 and λ_2 are the balancing parameters.

Experiments

Datasets

MIRFlickr-25K consists of 25000 image-text pairs that annotated with 1 or more of 24 categories from the Flickr website (Huiskes and Lew 2008). Following (Lin et al. 2015), in the experiment, we only selected the instances with frequent text tags that appearing more than 20 times. As such, 16738 pairs of them were remained, in which 2000 pairs of them were selected as the query set and rest of them were selected for the training and retrieval sets.

IAPR TC-12 consists of 20000 pairs of image-text instances annotated with 255 categories (Rasiwasia et al. 2010). For each pair of them, the image is represented as a 512-dimensional GIST vector, and the text is represented as a 2912-dimensional BoW vector. In the experiment, only 2000 pairs were randomly selected as the query set, and the rest were selected for the training and retrieval sets.

NUS-WIDE consists of 269648 pairs of image-text instances annotated with 81 categories (Chua et al. 2009). In the experiment, pairs of instances that annotated with 10 most frequent categories in the original 81 categories were chosen. Thus, there were 186577 labeled pairs were selected, in which only 2100 pairs were treated as query set and the rest were treated as the training and retrieval sets.

Evaluation Metrics

The performances of two tasks on cross-modal retrieval are evaluated in this section, which are text to image (T→I) and image to text (I→T) tasks, respectively. As a widely accepted metric for the performance evaluation in the cross-modal retrieval, mean average precision (mAP) is adopted as one of the evaluation metrics in this paper. mAP is actually the mean of average precision, which can verify the performance of the model more comprehensively and accurately (Wang et al. 2016). Moreover, the precision recall and top-N precision are also the widely accepted performance evaluation metrics for cross-modal retrieval tasks.

| Task | Method | MIRFlickr-25K | | | | IAPR TC-12 | | | | NUS-WIDE | | | |
|------|-------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | | 16 bits | 32 bits | 64 bits | 128 bits | 16 bits | 32 bits | 64 bits | 128 bits | 16 bits | 32 bits | 64 bits | 128 bits |
| I→T | DCMH | 0.7376 | 0.7466 | 0.7469 | 0.7547 | 0.4526 | 0.4732 | 0.4844 | 0.4983 | 0.5783 | 0.5921 | 0.5984 | 0.6097 |
| | SSAH | 0.7654 | 0.7728 | 0.7809 | 0.7892 | 0.5381 | 0.5668 | 0.5862 | 0.5992 | 0.6036 | 0.6187 | 0.6404 | 0.6428 |
| | UKD | 0.7138 | 0.7184 | 0.7255 | 0.7201 | 0.4812 | 0.4920 | 0.4941 | 0.5027 | 0.6149 | 0.6378 | 0.6386 | 0.6454 |
| | DCHUC | 0.7576 | 0.7557 | 0.7611 | 0.7495 | 0.5545 | 0.6023 | 0.6257 | 0.6484 | 0.7501 | 0.7759 | 0.7976 | 0.7953 |
| | DAEH | 0.7802 | 0.7916 | 0.7984 | 0.8027 | 0.5428 | 0.5796 | 0.6011 | 0.6087 | 0.7299 | 0.7534 | 0.7708 | 0.7769 |
| | SKDCH | 0.7216 | 0.7335 | 0.7382 | 0.7403 | 0.5421 | 0.5714 | 0.5998 | 0.6103 | 0.7314 | 0.7444 | 0.7642 | 0.7701 |
| | OLSH | 0.5784 | 0.5918 | 0.5981 | 0.5961 | 0.3465 | 0.3519 | 0.3619 | 0.3665 | 0.5312 | 0.5257 | 0.5295 | 0.5354 |
| | OCMFH | 0.5860 | 0.5881 | 0.5894 | 0.5875 | 0.3056 | 0.3024 | 0.2961 | 0.3005 | 0.3761 | 0.3778 | 0.3821 | 0.3866 |
| | LEMON | 0.7226 | 0.7408 | 0.7402 | 0.7437 | 0.4879 | 0.5101 | 0.5252 | 0.5318 | 0.6535 | 0.6600 | 0.6594 | 0.6702 |
| | SDOH | 0.7101 | 0.7268 | 0.7332 | 0.7330 | 0.4787 | 0.4905 | 0.5020 | 0.5084 | 0.6449 | 0.6582 | 0.6716 | 0.6780 |
| ROH | 0.7346 | 0.7525 | 0.7585 | 0.7626 | 0.4842 | 0.5094 | 0.5347 | 0.5471 | 0.6701 | 0.6854 | 0.6928 | 0.7001 | |
| | LCDH | 0.7898 | 0.7992 | 0.8034 | 0.8041 | 0.5656 | 0.6093 | 0.6301 | 0.6499 | 0.7564 | 0.7803 | 0.8007 | 0.8015 |
| T→I | DCMH | 0.7628 | 0.7733 | 0.7792 | 0.7857 | 0.5185 | 0.5378 | 0.5468 | 0.5596 | 0.6277 | 0.6413 | 0.6491 | 0.6503 |
| | SSAH | 0.7768 | 0.7847 | 0.7811 | 0.7823 | 0.5392 | 0.5646 | 0.5873 | 0.5998 | 0.6140 | 0.6285 | 0.6306 | 0.6324 |
| | UKD | 0.7156 | 0.7166 | 0.7214 | 0.7190 | 0.4886 | 0.5017 | 0.5145 | 0.5172 | 0.6304 | 0.6562 | 0.6573 | 0.6639 |
| | DCHUC | 0.7680 | 0.7791 | 0.7753 | 0.7681 | 0.5393 | 0.6039 | 0.6431 | 0.6734 | 0.7051 | 0.7243 | 0.7479 | 0.7453 |
| | DAEH | 0.7597 | 0.7667 | 0.7741 | 0.7813 | 0.5132 | 0.5558 | 0.5799 | 0.5905 | 0.7130 | 0.7353 | 0.7455 | 0.7501 |
| | SKDCH | 0.7814 | 0.7985 | 0.8067 | 0.7982 | 0.5404 | 0.6021 | 0.6398 | 0.6701 | 0.7228 | 0.7404 | 0.7519 | 0.7583 |
| | OLSH | 0.5988 | 0.5996 | 0.6011 | 0.6031 | 0.3521 | 0.3668 | 0.3625 | 0.3697 | 0.5604 | 0.5638 | 0.5880 | 0.5841 |
| | OCMFH | 0.5829 | 0.6094 | 0.6038 | 0.6024 | 0.3311 | 0.3173 | 0.3286 | 0.3203 | 0.4092 | 0.4175 | 0.4282 | 0.4330 |
| | LEMON | 0.8137 | 0.8270 | 0.8333 | 0.8304 | 0.5956 | 0.6185 | 0.6507 | 0.6687 | 0.7772 | 0.7850 | 0.8053 | 0.8018 |
| | SDOH | 0.7605 | 0.7961 | 0.7981 | 0.8032 | 0.5644 | 0.5918 | 0.6223 | 0.6441 | 0.7734 | 0.7859 | 0.7982 | 0.7991 |
| ROH | 0.8194 | 0.8362 | 0.8443 | 0.8481 | 0.5765 | 0.6147 | 0.6535 | 0.6795 | 0.7940 | 0.8092 | 0.8162 | 0.8264 | |
| | LCDH | 0.8223 | 0.8480 | 0.8556 | 0.8574 | 0.5987 | 0.6229 | 0.6574 | 0.6810 | 0.8018 | 0.8116 | 0.8268 | 0.8311 |

Table 1: mAP results of LCDH and baseline methods on MIRFlickr-25K, IAPR TC-12 and NUS-WIDE.

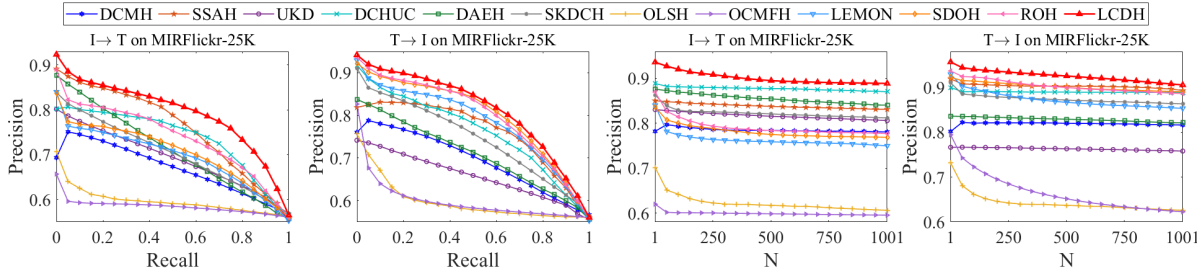


Figure 2: Precision-recall and Top-N precision curves of LCDH and baselines on MIRFlickr-25K with 128 bits hash codes.

Baseline Methods and Implementation Details

To verify the superior performances on cross-modal retrieval of LCDH, we compared LCDH with 6 offline cross-modal hashing methods (i.e., DCMH₂₀₁₇, AGAH₂₀₁₉, UKD₂₀₂₀, DCHUC₂₀₂₂, DAEH₂₀₂₂ and SKDCH₂₀₂₃), and 5 online cross-modal hashing methods (i.e., OLSH₂₀₁₉, OCMFH₂₀₂₀, LEMON₂₀₂₀, SDOH₂₀₂₂, ROH₂₀₂₃). Notably, UKD₂₀₂₀ and SKDCH₂₀₂₃ are the knowledge distillation-based hashing methods. All experiments were conducted on a workstation that running Windows 10 professional operation system with AMD Ryzen9 5900X CPU and NVIDIA GeForce RTX3090 GPU. The versions of CUDA is 11.3, of python is 3.9.8 and of pytorch is 1.10.1.

Results on Datasets

As shown in Table 1, the mAP results of the proposed method outperform other baseline methods in terms of I→T

and T→I tasks on MIRFlickr-25K, IAPR TC-12 and NUS-WIDE. Normally, the offline cross-modal hashing baseline methods will outperform the online ones. This is mainly because the offline methods learns hash codes and functions in two steps with full database, while the online methods learns hash codes and updates hash functions with data chunks of partial database. The results of the proposed LCDH and baseline methods on mAP, precision recall curve and top-N precision curve are as follows.

Results on mAP From Table 1, we can have the following observations and analyses of mAP results.

- LCDH outperforms knowledge distillation-based UKD and SKDCH methods, verifying that both knowledge distillation framework and features extraction modules are effective and have considerable contributions to the performance of LCDH. The main reason may be that the

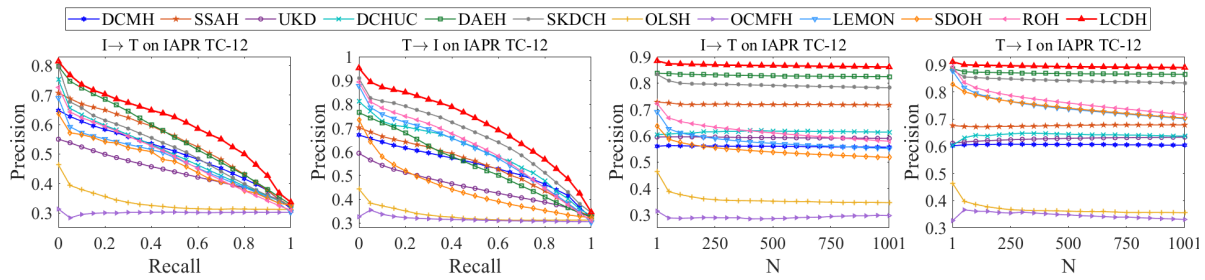


Figure 3: Precision-recall and Top-N precision curves of LCDH and baselines on IAPR TC-12 with 128 bits hash codes.

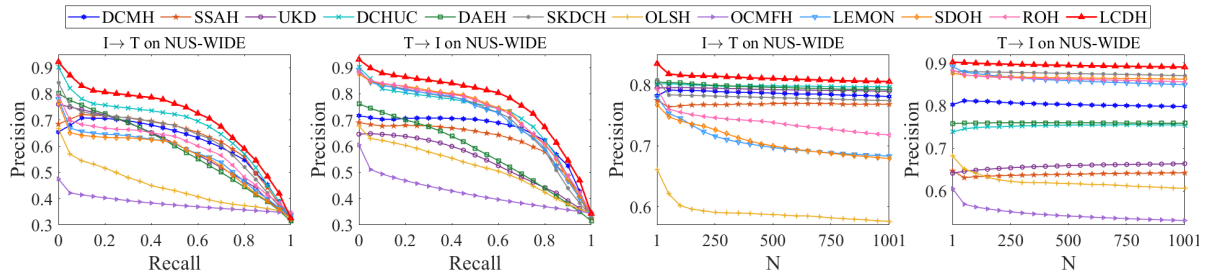


Figure 4: Precision-recall and Top-N precision curves of LCDH and baselines on NUS-WIDE with 128 bits hash codes.

feature extraction by CLIP and feature representation enhanced by the attention module in LCDH are more effective than that in UKD and SKDCH.

- All methods deliver outstanding results on mAP for MIRFlickr-25K and NUS-WIDE, but ordinary one for IAPR TC-12. The reason may be that, data in IAPR TC-12 is with much higher category number and feature dimensions, than that in MIRFlickr-25K and NUS-WIDE.
- As an online cross-modal hashing method, LCDH still can outperform some state-of-the-art offline cross-modal hashing methods in all cases. The main reason may be that, LCDH has powerful semantic preservation ability by fusing contrastive features extracted by CLIP, and the fused feature is further fed into the attention module to enhance the feature representation.
- LCDH has achieved obvious improvements on mAP results for both I→T and T→I tasks, especially when the length of hash codes is short. Notably, LCDH can deliver the remarkable and best results on mAP, of both I→T and T→I tasks for all three datasets.

Results on Precision Recall Curve By measuring the relevance of the precision and retrieval results, precision recall curve can reveal the relationship between them. Therefore, the higher the precision recall curve, the better the performance on retrieval. From Figs. 2, 3 and 4, we have the following observations and analyses of precision recall curve.

- In all case, there are obvious gaps between the curves of the proposed LCDH and the second best baseline methods. This is consistent to the mAP results, indicating that LCDH can achieve competitive retrieval performance on precision in all levels of recall.
- All the methods including the proposed LCDH perform

better on T→I than I→T tasks, in terms of the precision recall curve. The main reason may be that, the fusion of contrastive learning on cross-modal features can well extract the coexistent semantic relevance representation, which is further enhanced by the attention module for narrowing the heterogeneous gaps on semantic.

- In all levels of the curves, LCDH can deliver the highest precision consistently, for both I→T and T→I tasks on all three datasets. This indicates that LCDH is able to effectively capture the adequate coexistent semantic relevance, by the well designed feature extraction model and knowledge distillation framework in LCDH.

Results on Top-N Precision Curve Normally, top-N precision curve can reveal the proportion of the top N results, which are with the highest probability in prediction results containing correct label. Therefore, the higher the top-N precision curve, the better the performance on retrieval. From Figs. 2, 3 and 4, we can have the following observations and analyses of top-N precision curve.

- There exists significant gaps between the top-N precision curves of the proposed LCDH and other baseline methods, especially on MIRFlickr-25K and IAPR TC-12.
- LCDH surpasses other baseline methods with a diverse numbers of retrieval instances, in terms of top-N precision on all three datasets.
- When the retrieval number of instance increasing largely, the proposed LCDH can still achieve the prominent and stable retrieval performance. This verifies that, LCDH can effectively and correctly transfer the coexistent semantic relevance information to student network by the similarity approximation and knowledge distillation framework proposed in LCDH.

| Task | Variants | MIRFlickr-25K | | | | IAPR TC-12 | | | | NUS-WIDE | | | |
|------|--------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | | 16 bits | 32 bits | 64 bits | 128 bits | 16 bits | 32 bits | 64 bits | 128 bits | 16 bits | 32 bits | 64 bits | 128 bits |
| I→T | <i>Log-Similarity</i> – | 0.7785 | 0.7813 | 0.7907 | 0.7981 | 0.5524 | 0.5838 | 0.6074 | 0.6253 | 0.7462 | 0.7697 | 0.7915 | 0.7991 |
| | <i>Offline Hashing</i> – | 0.7609 | 0.7664 | 0.7699 | 0.7702 | 0.5389 | 0.5573 | 0.5897 | 0.6004 | 0.7277 | 0.7439 | 0.7741 | 0.7865 |
| | <i>CLIP</i> – | 0.7671 | 0.7698 | 0.7710 | 0.7716 | 0.5427 | 0.5684 | 0.5913 | 0.6107 | 0.7324 | 0.7572 | 0.7890 | 0.7906 |
| | LCDH | 0.7898 | 0.7992 | 0.8034 | 0.8041 | 0.5656 | 0.6093 | 0.6301 | 0.6499 | 0.7564 | 0.7803 | 0.8007 | 0.8015 |
| T→I | <i>Log-Similarity</i> – | 0.8069 | 0.8257 | 0.8447 | 0.8532 | 0.5854 | 0.6043 | 0.6260 | 0.6598 | 0.7875 | 0.8011 | 0.8109 | 0.8212 |
| | <i>Offline Hashing</i> – | 0.7943 | 0.8102 | 0.8228 | 0.8395 | 0.5753 | 0.5874 | 0.6183 | 0.6444 | 0.7790 | 0.7906 | 0.7975 | 0.8090 |
| | <i>CLIP</i> – | 0.7998 | 0.8144 | 0.8316 | 0.8469 | 0.5791 | 0.5990 | 0.6247 | 0.6585 | 0.7816 | 0.7951 | 0.8002 | 0.8133 |
| | LCDH | 0.8223 | 0.8480 | 0.8556 | 0.8574 | 0.5987 | 0.6229 | 0.6574 | 0.6810 | 0.8018 | 0.8116 | 0.8268 | 0.8311 |

Table 2: The ablation results of LCDH on the three datasets with various code lengths.

Ablation Analysis

mAP results of the ablation analysis are represented in Table 2, where ‘–’ means the corresponding module in the proposed LCDH is dropped and the optimization will be redid.

As shown in Table 2, ‘*Offline Hashing*’ and ‘*CLIP*’ have considerable contributions to performances of LCDH, while ‘*Log-Similarity*’ have relatively weak contribution to the performance of LCDH. The results of ‘*Offline Hashing*’ and ‘*CLIP*’ have verified the effectiveness of knowledge distillation framework and the CLIP-based feature extraction in the proposed LCDH. That is, the deep semantic information is well extracted and preserved in teacher network, and then effectively transferred to the student network for guiding the online cross-modal hashing. Moreover, the results also demonstrate that log-likelihood similarity preservation term is well designed to guide the training in student network. Finally, the LCDH with full modules had achieved the best results in all cases, revealing that all the modules have indeed contributions to the performance of the proposed LCDH.

Convergence Analysis

By recording and illustrating the values of the objective function (i.e., Eq. (11)), the convergence analysis for LCDH is provided as shown in Fig. 5. From the figure, it can be observed easily that the losses (the values of the objective function) of LCDH are reducing quickly and becoming stable after around ten iterations for all datasets. This verifies that LCDH can quickly become convergent and deliver a stable and consistent performance. That is, the quick convergence of losses can prove that LCDH effectively transfers the co-existent semantic relevance on modalities to guide the online hashing by the knowledge distillation framework.

Parameter Sensitivity

In the objective function of LCDH, parameter λ_1 controls the impact of the similarity preservation in teacher network, while parameter λ_2 controls the impact of similarity approximation for offline and online similarity in knowledge distillation. The ranges of λ_1 and λ_2 are set to $\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1e^0, 1e^1, 1e^2, 1e^3, 1e^4, 1e^5\}$. As shown in Fig. 6(a) and Fig. 6(b), LCDH achieved the best performance on MIRFlickr-25K for both I→T and T→I retrieval tasks when $\lambda_1 = 1e^4$, $\lambda_2 = 1e^0$. From the figures, we also can find that both λ_1 and λ_2 have make

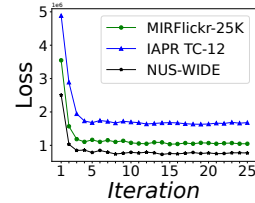
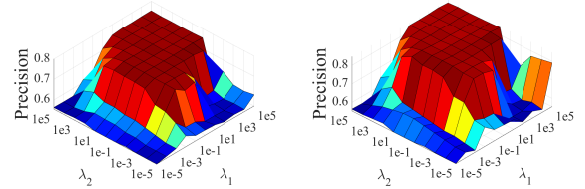


Figure 5: The convergence analysis for LCDH on MIRFlickr-25K, IAPR TC-12 and NUS-WIDE @ 128 bits.



(a) Parameter sensitivity of λ_1 and λ_2 for I→T tasks on MIRFlickr-25K (b) Parameter sensitivity of λ_1 and λ_2 for T→I tasks on MIRFlickr-25K

Figure 6: The parameter analysis for LCDH of λ_1 and λ_2 , for I→T and T→I tasks on MIRFlickr-25K @ 128 bits.

contributions to the performances of LCDH to great extents. It indicates that the deep feature extraction, representation enhancement, similarity approximation and the knowledge distillation framework are effectively designed in LCDH.

Conclusion

This paper proposes a lightweight contrastive distilled hashing (LCDH) method for online cross-modal retrieval. LCDH fuses CLIP-extracted cross-modal features and enhances representations using an attention module. Binary codes with label information are generated via a fully connected layer to align offline and online similarity matrices. In the student network, image and text features from VGG16 and BoW are similarly transformed into binary codes. By approximating similarities and leveraging semantic relevance from the teacher network, LCDH effectively bridges offline and online hashing, achieving competitive retrieval performance with state-of-the-art offline methods, while maintaining a lightweight structure.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62302112, Grant 62006048, and Grant 62176065; in part by the Guangdong Pearl River Talent Program under Grant 2023QN10X503; in part by the Guangzhou Basic and Applied Basic Research Foundation under Grant 2025A04J3378; in part by the Guangdong Provincial National Science Foundation under Grant 2021A1515012017; and in part by the Guangzhou University Research Project under Grant RQ2021013.

References

- Chen, Y.; Chen, R.; Li, Q.; Fang, X.; Li, J.; and Wong, W. K. 2024. Denoising High-Order Graph Clustering. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 3111–3124. IEEE.
- Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and 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, CIVR '09*. New York, NY, USA: ACM. ISBN 9781605584805.
- Cui, C.; Ren, Y.; Pu, J.; Li, J.; Pu, X.; Wu, T.; Shi, Y.; and He, L. 2023a. A Novel Approach for Effective Multi-View Clustering with Information-Theoretic Perspective. In *Advances in Neural Information Processing Systems*, volume 36, 44847–44859.
- Cui, C.; Ren, Y.; Pu, J.; Pu, X.; and He, L. 2023b. Deep Multi-view Subspace Clustering with Anchor Graph. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 3577–3585. Main Track.
- Hu, H.; Xie, L.; Hong, R.; and Tian, Q. 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3123–3132.
- Huang, J.; Kang, P.; Han, N.; Chen, Y.; Fang, X.; Gao, H.; and Zhou, G. 2023. Two-stage asymmetric similarity preserving hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 429–444.
- Huiskes, M. J.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 39–43. New York, NY, USA: ACM. ISBN 9781605583129.
- Jiang, K.; Wong, W. K.; Fang, X.; Li, J.; Qin, J.; and Xie, S. 2023a. Random Online Hashing for Cross-Modal Retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. Doi: 10.1109/TNNLS.2023.3330975.
- Jiang, L.; Fang, X.; Sun, W.; Han, N.; and Teng, S. 2023b. Low-rank constraint based dual projections learning for dimensionality reduction. *Signal Process.*, 204: 108817.
- Jiang, L.; Wu, J.; Zhao, S.; and Li, J. 2024a. Coding self-representative and label-relaxed hashing for cross-modal retrieval. *Pattern Recognition Letters*, 185: 264–270.
- Jiang, L.; Wu, J.; Zhao, S.; Li, J.; and Ma, S. 2024b. Joint Category Compactness and Disturbance Reduction for Cross-Domain Classification. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–14. Doi: 10.1109/TIM.2024.3368421.
- Jiang, Q.-Y.; and Li, W.-J. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3232–3240.
- Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; and Tao, D. 2018. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4242–4251.
- Li, J.; Wong, W. K.; Jiang, L.; Fang, X.; Xie, S.; and Xu, Y. 2024. CKDH: CLIP-based Knowledge Distillation Hashing for Cross-modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6530–6541.
- Lin, M.; Ji, R.; Liu, H.; and Wu, Y. 2018. Supervised online hashing via hadamard codebook learning. In *Proceedings of the 26th ACM international conference on Multimedia*, 1635–1643.
- Lin, Z.; Ding, G.; Hu, M.; and Wang, J. 2015. Semantics-preserving hashing for cross-view retrieval. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3864–3872.
- Ling, Y.; Chen, J.; Ren, Y.; Pu, X.; Xu, J.; Zhu, X.; and He, L. 2023. Dual label-guided graph refinement for multi-view graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8791–8798.
- Liu, J.; Fei, L.; Zhao, S.; Wen, J.; Rida, I.; and Xu, Y. 2022. Supervised Discrete Online Hashing For Large-scale Cross-modal Retrieval. In *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 896–902.
- Liu, W.; Wang, J.; Ji, R.; Jiang, Y.-G.; and Chang, S.-F. 2012. Supervised hashing with kernels. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2074–2081. Doi: 10.1109/CVPR.2012.6247912.
- Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, 251–260. New York, NY, USA: ACM. ISBN 9781605589336.
- Shi, Y.; Zhao, Y.; Liu, X.; Zheng, F.; Ou, W.; You, X.; and Peng, Q. 2022. Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 7255–7268.
- Su, M.; Gu, G.; Ren, X.; Fu, H.; and Zhao, Y. 2023. Semi-Supervised Knowledge Distillation for Cross-Modal Hashing. *IEEE Transactions on Multimedia*, 25: 662–675.
- Tu, R.-C.; Mao, X.-L.; Ma, B.; Hu, Y.; Yan, T.; Wei, W.; and Huang, H. 2022. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(2): 560–572.

- Wang, D.; Wang, Q.; An, Y.; Gao, X.; and Tian, Y. 2020. On-line collective matrix factorization hashing for large-scale cross-media retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 1409–1418.
- Wang, K.; Yin, Q.; Wang, W.; Wu, S.; and Wang, L. 2016. A Comprehensive Survey on Cross-modal Retrieval. arXiv:1607.06215.
- Wang, Y.; Luo, X.; and Xu, X.-S. 2020. Label Embedding Online Hashing for Cross-Modal Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*, 871–879. New York, NY, USA: ACM. ISBN 9781450379885.
- Xu, J.; Ren, Y.; Shi, X.; Shen, H. T.; and Zhu, X. 2023. UNTIE: Clustering analysis with disentanglement in multi-view information fusion. *Information Fusion*, 100: 101937.
- Yao, T.; Wang, G.; Yan, L.; Kong, X.; Su, Q.; Zhang, C.; and Tian, Q. 2019. Online latent semantic hashing for cross-media retrieval. *Pattern Recognition*, 89: 1–11.
- Zhang, D.; and Li, W. 2014. Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization. In Brodley, C. E.; and Stone, P., eds., *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, 2177–2183. AAAI Press.
- Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Chen, H.; Dou, Z.; and Wen, J.-R. 2024. Large Language Models for Information Retrieval: A Survey. arXiv:2308.07107.