Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 7992047, 16 pages
https://doi.org/10.1155/2023/7992047

Research Article

CCAH: A CLIP-Based Cycle Alignment Hashing Method for Unsupervised Vision-Text Retrieval

Mingyong Li, Longfei Ma, Yewen Li, and Mingyuan Ge

College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China

Correspondence should be addressed to Mingyong Li; limingyong@cqnu.edu.cn

Received 13 November 2022; Revised 5 December 2022; Accepted 6 December 2022; Published 23 February 2023

Academic Editor: Gennaro Vessio

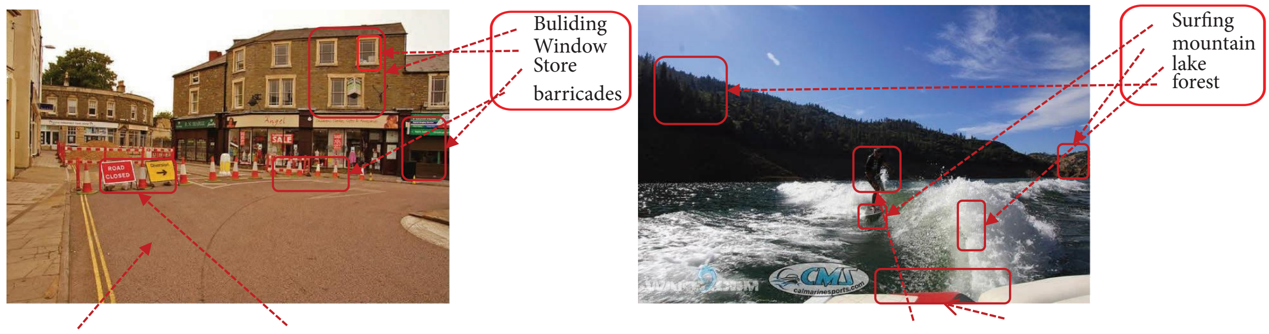
Due to the advantages of low storage cost and fast retrieval efficiency, deep hashing methods are widely used in cross-modal retrieval. Images are usually accompanied by corresponding text descriptions rather than labels. Therefore, unsupervised methods have been widely concerned. However, due to the modal divide and semantic differences, existing unsupervised methods cannot adequately bridge the modal differences, leading to suboptimal retrieval results. In this paper, we propose CLIP-based cycle alignment hashing for unsupervised vision-text retrieval (CCAH), which aims to exploit the semantic link between the original features of modalities and the reconstructed features. Firstly, we design a modal cyclic interaction method that aligns semantically within intramodality, where one modal feature reconstructs another modal feature, thus taking full account of the semantic similarity between intramodal and intermodal relationships. Secondly, introducing GAT into cross-modal retrieval tasks. We consider the influence of text neighbour nodes and add attention mechanisms to capture the global features of text modalities. Thirdly, Fine-grained extraction of image features using the CLIP visual coder. Finally, hash encoding is learned through hash functions. The experiments demonstrate on three widely used datasets that our proposed CCAH achieves satisfactory results in total retrieval accuracy. Our code can be found at: https://github.com/CQYIO/CCAH.git.

1. Introduction

As the internet and social networking grow rapidly, multimedia information data such as images and texts are increasing dramatically, and it is a great challenge to retrieve these data efficiently. Cross-modal retrieval aims to search for heterogeneous modal data with a similar semantic representation by one modality. Hashing methods[1–8] are widely used in retrieval tasks to improve storage and computational efficiency. Cross-modal hashing methods attempt to represent heterogeneous modal data as compact binary codes while maintaining semantic similarity between different modal data in a common hidden space.

Cross-modal hashing methods fall into two broad categories: supervised methods and unsupervised methods. Commonly available supervised hashing methods [2, 7, 9–13] have demonstrated significant performance. The principle is to use hand-labeled label information or precomputed similarity matrices to guide model training and learning of binary codes. Unfortunately, in real-world and more challenging scenarios, images are often accompanied by their textual description, but difficult to obtain their labels, categories, or tags.

Recently, an increasing number of research hotspots have emerged in unsupervised cross-modal hashing methods. Unsupervised hashing methods [1, 14–18] attempt to get rid of the model's reliance on manually annotated data during training, relying solely on the features of the data itself, and demonstrate superior performance. However, a common drawback of the above-unsupervised approach is that the co-occurrence information inherent in the vision-text is easily overlooked in the high-level semantic feature extraction process due to the lack of labeling information guidance (Figure 1). This further leads to unsupervised models that are unable to accurately capture the semantic connections between different modal data, making retrieval accuracy suboptimal. In view of this, we point out that hash codes of images and text that appear in pairs should have



A road closed and detour sign at the end of a road.

A somewhat distant man is riding behind the boat on his board.

Figure 1: As images contain richer higher-order semantic information than text, text retrieval of images usually pays attention to image regions that are consistent with the text representation, resulting in missing vision modal semantics and reduced accuracy of text retrieved images.

either a minimum Hamming distance or a maximum degree of semantic similarity.

In addition, most existing cross-modal methods focus on the alignment of semantic features between cross-modal data (GAN [19]). Simplifying the semantic association of reconstructed features within a modality with the original features makes the generated hash codes not perfectly compatible with cross-modal retrieval tasks. Inevitably, there is an inherent modal divide problem in high-level semantic interactions, where one cannot pay attention to both intramodal and intermodal semantic information of one's modality, nor can one bridge the alignment of modal features and hash encoding, resulting in retrieval results that do not achieve optimal solutions.

To solve the above problem, in this paper we propose a novel deep unsupervised cyclic semantic alignment cross-modal hashing method termed CLIP-based cycle alignment hashing for unsupervised vision-text retrieval (CCAH). CCAH is an end-to-end learning framework that simultaneously notices both intramodal and intermodal semantic features and hash code consistency. Our CCAH network model consists of three components: deep feature extraction, cycle alignment, and hash encoding learning. Previous unsupervised network models have suffered from a problem of low accuracy in text retrieval images. It is well known that in image text pairs, images contain richer semantic information and can extract higher-level semantic representations at a finer level. Compared to the corresponding text description, (e.g.,: BOW) the text has relatively little semantic information, and often only a few keywords can be matched to the described image area (attention points). Moreover, the text has a contextual relationship, and the same word may represent different semantic information, resulting in text retrieval images that are often less accurate than image retrieval text. We propose to consider the text as data in a graph structure, transforming text features into node information in the graph, further fusing sparse text features by using GAT networks, and fusing related neighboring node information with the original nodes in an attention scoring mechanism, while the attention score indicates the closeness of the connection between different nodes, with higher scores being more closely related. And the auto-encoder is used to encode and decode the extracted modal features. Our contribution to this work is as follows:

(i) We propose a new deep hash network model called CCAH. CLIP is used as a visual coder to extract fine-grained features. The GAT network is also used for feature extraction of text modalities.

(ii) We propose a circular alignment method to align image features with features extracted by auto-encoder, and then align the features after mapping them to the text modality space to ensure semantic links between modalities and vice versa.

(iii) The experiments demonstrate that our model achieves satisfactory results in terms of final total retrieval accuracy under three commonly used multimodal datasets.

2. Related Work

Currently, cross-modal hash retrieval is broadly divided into supervised and unsupervised hashing. Supervised hashing methods have better performance compared to unsupervised methods with the aid of labels or similarity matrices to avoid redundant information interference.

2.1. Supervised Hashing Methods.

Supervised hashing methods: supervised hashing methods use manually annotated label information or load predefined similarity matrices to guide the training of binary encoding between different modalities and have shown excellent implementation in multimodal data retrieval. Recently, many supervised hashing methods have been used to continuously improve the retrieval accuracy benchmarks. TDH [20] uses triples to flexibly capture a variety of higher-level similarities, rather than the simple similarity or dissimilarity of binary groups, sorting to optimize intraclass and interclass variation; SCM [13] learns the hash function bit by bit using supervised information in linear time complexity; DOH [21] learns ordinal representations to generate ranking-based hash codes by leveraging the ranking structure of feature space from both local and global views; Seph [3] uses a probability distribution, which is approximated by minimizing Kullback–Leibler divergence, to a hash code learned in Hamming space; QCH [9] proposes to simplify the optimization process by transforming the multimodal objective function into a unimodal formalism; MCSCH [12] proposed a multiscale association mining strategy, which is a multiscale feature-guided sequence hashing method; DLFH [11] introduces a discrete learning algorithm that learns binary hash codes directly, without the need for successive relaxations. However, the above methods require a lot of manual and financial effort to label the dataset during the hash function learning process, which is often unrealistic in real-life scenarios. And without labeling information, the retrieval accuracy inevitably degrades.

2.2. Unsupervised Hashing Methods. Unsupervised hashing methods: to reduce the need for manual annotation information during model training, unsupervised cross-modal hashing methods are proposed. CVH [1] learns binary codes by minimizing the similarity-weighted Hamming distance; IMH [6] builds two intramodal similarity matrices based on neighbor relations; CMFH [16] uses matrix decomposition to address the semantic relevance of different modalities and maps heterogeneous modal data into a hidden state space; UDCMH-based [17] learning of features and hash codes under Laplacian and discrete constraints; DJSRH [14] fuses semantic information into the affinity matrix to calculate potential correlations between modes; DSAH [22] aligns intramodel and intermodal data by fusing them using semantic similarity alignment and heterogeneous modal data reconstruction; JIMFH [23] combines intramodal and intermodal hash codes to obtain the final hash code; DBRC [24] proposes a framework with adaptive binary reconstruction that allows discrete hash codes to be learned directly; HNH [25] weighted the original similarities using Hadamard products and created a joint similarity matrix using linear combinations. Although these unsupervised cross-modal hash models have achieved better results regarding the colinear information of image text pairs, they still ignore part of the image information, resulting in poor accuracy of text retrieved images.

3. Problem Formulation

3.1. Problem Definition. Suppose we have m image text pairs, We define our data structure as $O = \{o_i\}_{i=1}^{m}$. We define I_i to represent the i-th image and T_j to represent the j-th text. Each image text pair instance can be represented as $o_k = \{I_k, T_k\}$. We define the representation of the feature dimension as F. The semantic features extracted by the visual feature encoder denoted as F_I and $F_I \in R^{m\times D_I}$, D_I is the high-level dimensional feature representation of the image obtained by passing the original vision through the image encoder. We also define the feature representation of the text after the text encoder as $F_T \in R^{m\times D_T}$, D_T denotes the high-level feature dimensional representation of the text, and m is the number of sample instance points. In addition, we define the hash code representation as $B_* \in \{-1, +1\}^{m\times c}$, and $* \in \{I, T\}$. Here c denotes the length of the hash code, and the hash code of the i-th original data in B_* is denoted $b_{*,i}$. In addition we define the cosine similarity loss function for paired image text as $\cos(\cdot)$, and use the $\operatorname{sign}(\cdot)$ function for element wise symbolic functions. The definitions are as follows:

$$\cos(a, b) = \frac{ab}{\|a\|\|b\|},$$
$$\operatorname{sign}(x) = \begin{cases} +1, & x \geq 0, \\ -1, & x < 0, \end{cases} \tag{1}$$

here we define $\|*\|$ to denote the l_2 regularization paradigm for the Frobenius regularization of vectors and matrices.

3.2. Model. In Figure 2, we show all the components of our model. The CLIP-based cycle alignment hashing for unsupervised vision-text retrieval (CCAH) consists of three parts, namely, the feature extraction part, the cycle semantic alignment part, and the hash coding learning part.

Graph networks [26] represent node information as a graph, transforming the graph topology into a constructed adjacency matrix by aggregating node-to-node associations, fusing the information of each node and its neighbors into a new node. With attention [27] showing advanced execution in NLP and CV, the attention mechanism is introduced into graph networks, where instead of just doing a simple fusion, the attention algorithm gives each node an attention score, and then fuses the different nodes for information. Less relevant feature words have a lower score, and feature words that are more relevant to them can receive a high attention score. In fusing this information, the influence of different feature words on the nodes is reinforced and better semantic information can be extracted.

Since our text is a 1386-dimensional feature vector representation, we treat these features as node data and each text can be represented as $f_i \in R^{1386}$. To obtain sufficient expressive power to transform the input features into higher-level features, after a learnable weighting matrix $W \in R^{f\times f}$ transformation, then self-attention is applied to the node.

$$e_{ij} = a\left(W\overrightarrow{f_i}, W\overrightarrow{f_j}\right), \tag{2}$$

where a is the attention calculation factor, e_{ij} denotes the importance of node j to node i. We calculate each neighboring node of node i. To make the coefficients easily comparable between different nodes, we use the softmax function to normalize all neighboring nodes.

$$a_{ij} = \operatorname{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(\operatorname{LeakyReLU}\left(\overrightarrow{a}^T\left[W\overrightarrow{h_i}\middle\|W\overrightarrow{h_j}\right]\right)\right)}{\sum_{k=1}^{m}\exp\left(\operatorname{LeakyReLU}\left(\overrightarrow{a}^T\left[W\overrightarrow{h_i}\middle\|W\overrightarrow{h_k}\right]\right)\right)}. \tag{3}$$

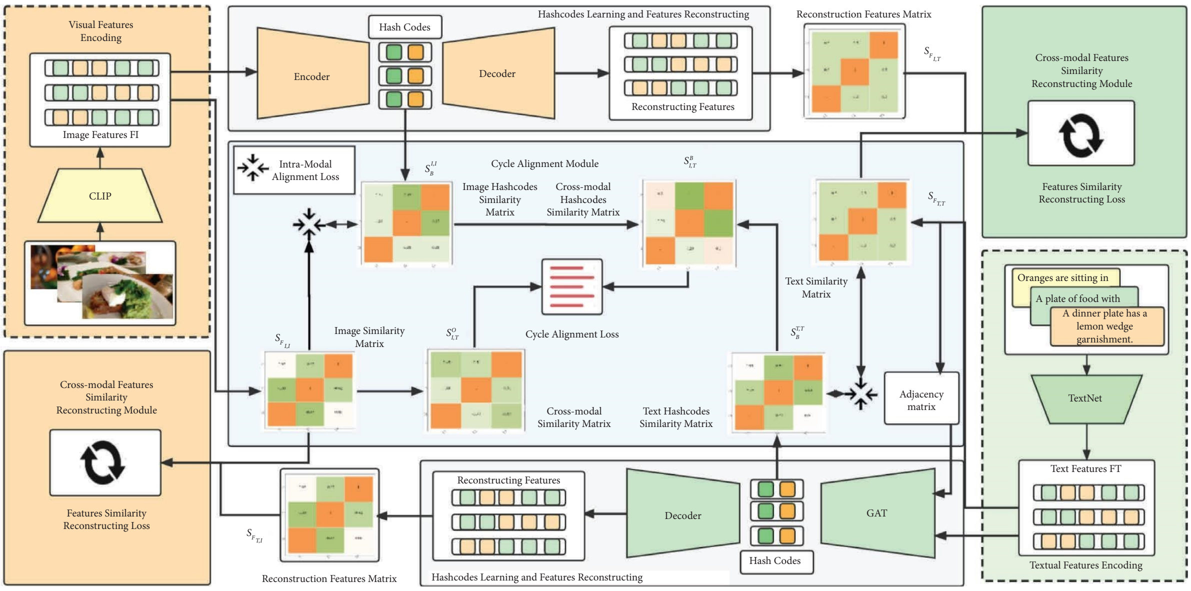

Visual Features Encoding
Image Features FI
CLIP
Hash Codes
Encoder
Decoder
Hashcodes Learning and Features Reconstructing
Reconstructing Features
Reconstruction Features Matrix
$S_{F_{I,T}}$
Cross-modal Features Similarity Reconstructing Module
Features Similarity Reconstructing Loss
Intra-Modal Alignment Loss
$S_B^{I,I}$
Cycle Alignment Module
Image Hashcodes Similarity Matrix
Cross-modal Hashcodes Similarity Matrix
$S_{I,T}^B$
$S_{F_{T,T}}$
Text Similarity Matrix
Cycle Alignment Loss
$S_{F_{I,I}}$
Image Similarity Matrix
$S_{I,T}^O$
$S_B^{T,T}$
Cross-modal Similarity Matrix
Text Hashcodes Similarity Matrix
Adjacency matrix
Oranges are sitting in
A plate of food with
A dinner plate has a lemon wedge garnishment.
TextNet
Text Features FT
Textual Features Encoding
Cross-modal Features Similarity Reconstructing Module
Features Similarity Reconstructing Loss
$S_{F_{T,I}}$
Reconstruction Features Matrix
Reconstructing Features
Decoder
GAT
Hash Codes
Hashcodes Learning and Features Reconstructing

Figure 2: The entire architecture of our model is represented in the figure above, with the orange region indicating the imaging modality and the green region is the text modality. We construct similarity matrices within and across modalities, and the generated hash matrices are also aligned between modalities to be able to guarantee semantic alignment within modalities, hash encoding, and features across modalities, and hash matrix to hash matrix alignment.

T represents the transpose of a vector. By doing this for all nodes, the node information of the adjacency matrix has been transformed into a new node vector containing the attentional features of each neighboring node, which is the most easily weighted fusion of semantic information that is lacking in the text modality, leading to a more powerful modal representation of the text modality. The graph attention network is a fusion of feature words associated with a certain feature word with its associated feature words using attention. And the weighted fusion employing the attention mechanism can obtain a new semantic feature representation containing the information of neighboring nodes.

3.2.1. Deep Feature Extraction. In order to extract richer information about the high-level semantic representations of the modalities, we design different modal encoders for different modal data. Since image modalities contain richer semantic information than text modalities, and the single-stream model (eg: ViLT [28]) cannot bridge the inherent modal gap across modalities, cannot perform optimal feature extraction for each modality, and has limited ability to mine semantic consistency information for heterogeneous data, we adopt a dual-stream model to extract semantic features for different modal data information and show excellent results throughout the training phase. The results were excellent throughout the training phase.

(1) Image Feature Extraction. CLIP [29] used a training method of contrast learning in unsupervised learning, using a dataset of huge size for training, compared to ViT [30], which yielded good quality results on several datasets. Using the CLIP pretrained model as a feature extractor for image modalities in our model. In the image section using CLIP's image encoder (encode-image), we feed the original image into the CLIP image encoder (Figure 3), and after extraction, we obtain a 1024-dimensional high-level semantic vector, which we define as $F_I \in R^{m\times 1024}$.

(2) Text Feature Extraction. We consider text modal data as not containing as much high-level semantic information as image data, but text semantics are contextually relevant. We treat the features of text as nodes of a graph and use graph attention networks (GAT [31]) to extract aggregated semantic information from text. GAT treats text features as nodes, and converts input features into higher-level features to obtain more powerful expression, introduces an attention mechanism, performs self-attention on nodes, and finds the attention weight coefficients between nodes; and by weighted summation of surrounding neighboring nodes, you can get information that aggregates all surrounding nodes, making the connection of text information more realistic (Figure 4). The text features are constructed as adjacency matrices, and the information in the adjacency matrices represents the linkage of text modalities, and the semantic representation of text can be better processed by weighting the features. The original text message is characterized by $F_T \in R^{m\times 1386}$.

For simplicity, we define the feature extractor as F. The mathematical notation of each modal feature extractor is defined as follows:

$$F_I = F\left(I;\Theta_I\right) F_T = F\left(T;\Theta_T\right), \tag{4}$$

where I and T are the original image and text. Θ_I and Θ_T are the parameters of the feature extractor. To this end, we can extract semantically rich high-level representation features for each modality, which can be used to fully explore the semantic relationships between the data and further guide modal alignment and hash code learning.

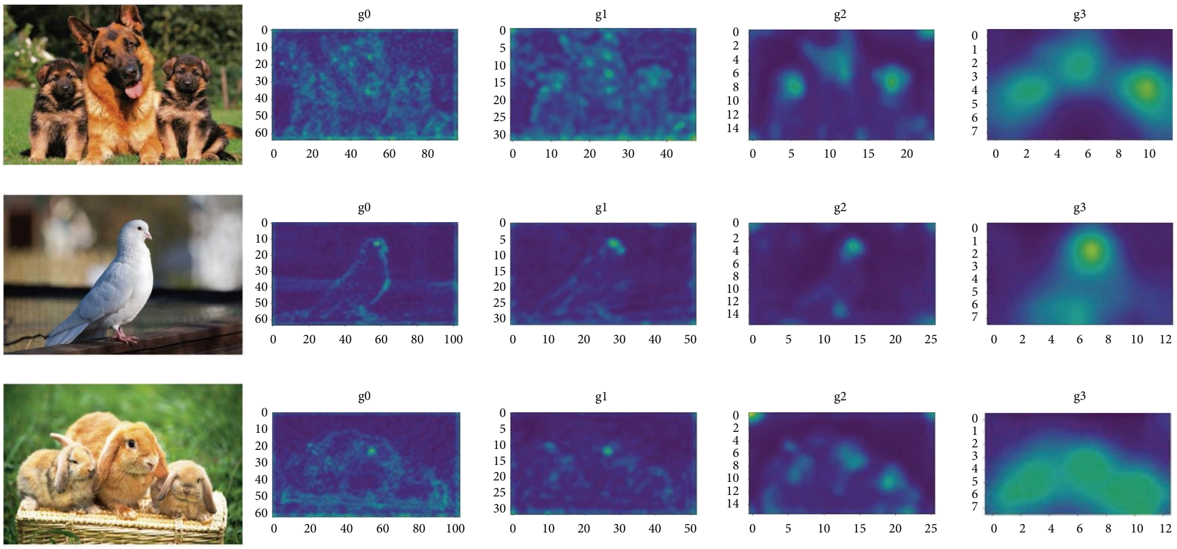
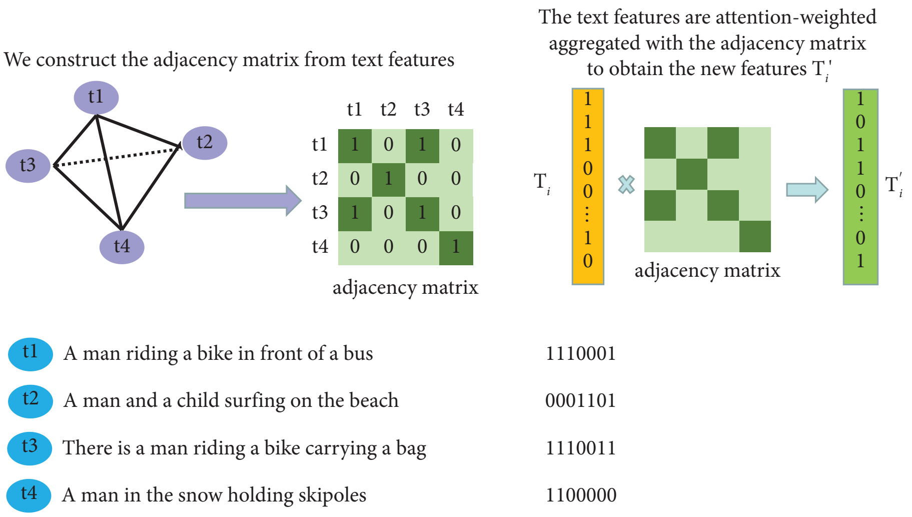


Figure 3: We use CLIP image encoder for the images, with the left side representing the original image and the right side representing the results of attention visualization for different levels of features.


We construct the adjacency matrix from text features
t1
t2
t3
t4
adjacency matrix
The text features are attention-weighted aggregated with the adjacency matrix to obtain the new features T$_i'$
T$_i$
1 1 1 0 0 ⋮ 1 0
adjacency matrix
T$_i'$
1 0 1 1 0 ⋮ 0 1

	t1	t2	t3	t4
t1	1	0	1	0
t2	0	1	0	0
t3	1	0	1	0
t4	0	0	0	1

t1	A man riding a bike in front of a bus	1110001
t2	A man and a child surfing on the beach	0001101
t3	There is a man riding a bike carrying a bag	1110011
t4	A man in the snow holding skipoles	1100000

Figure 4: We use the text features to construct the adjacency matrix, and then the text features are attention-weighted and summed with the constructed adjacency matrix to obtain the new text representation features, which have clustered the information of the surrounding neighbor nodes.

3.2.2. Cycle Alignment. To facilitate intramodal semantic feature alignment and to maintain cross-modal data semantic interaction, we propose the use of circular semantic alignment methods. The distance between semantically similar vision-texts is promoted to be close in the common representation space and vice versa makes the distance in the common representation space farther. To further align text and images we use intramodal and intermodal loss measures. We use auto-encode to compress the high-level semantic features into low-level semantic representations and to reconstruct this underlying semantic feature back into a feature of heterogeneous data. We define the function that compresses the high-level semantic representation as follows:

$$\begin{aligned} TV_I &= \mathrm{Enc}\left(F_I;\delta_I\right) TV_I \in R^{m\times c},\\ TV_T &= \mathrm{Enc}\left(F_T;\delta_T\right) TV_T \in R^{m\times c}, \end{aligned} \tag{5}$$

where F_* denotes the original features of the image and text, δ_* is defined as the parameter under each modal pass $\mathrm{Enc}(*)$, and $* \in \{I, T\}$.

The high-level semantic features extracted by the feature extractor are encoded and compressed by the encoder to obtain a true-value semantics with strong representational power and containing highly semantic features, which we then reconstruct back into a representation of the heterogeneous data by means of a decoder, which we define as follows:

$$F_I' = \text{Dec}\left(TV_T; \gamma_T\right) F_I' \in R^{m\times D_I},$$
$$F_T' = \text{Dec}\left(TV_I; \gamma_I\right) F_T' \in R^{m\times D_T}. \tag{6}$$

We input the features of the image (text) into the decoder and the semantic information obtained is then mapped to the feature space of the text (image) by the decoder to achieve semantic alignment between modalities. After obtaining the reconstructed features of heterogeneous data, to facilitate cross-modal information interaction. We semantically align the original image features with the text reconstructed by the decoder. To ensure that the resulting compressed feature vector represents the original high-dimensional feature representation, we align the high-dimensional features with the encoded features once as well, achieving intramodal semantic alignment.

(1) Intermodal. To facilitate information interaction between different data and achieve cross-modal semantic interaction, we use the semantic features obtained by the feature extractor of one modality to be mapped to the corresponding semantic space of another modality after being decoded by the auto-encoder. $F_{T,I}$ represents the vector representation after mapping the text features to the image feature space, and $F_{I,T}$ represents the vector representation obtained by mapping the image features to the text feature space. We construct the cross-modal semantic feature matrix $S_{F_{T,I}}$ and $S_{F_{I,T}}$. Alignment of different modal types is achieved by minimizing cross-modal semantic losses, with the following loss function:

$$L_{C-\text{inter1}} = \left\| S_{F_I'} - S_{F_I} \right\|^2,$$
$$L_{C-\text{inter2}} = \left\| S_{F_T'} - S_{F_T} \right\|^2. \tag{7}$$

The total intermodal loss is expressed as follows:

$$L_{C-\text{inter}} = L_{C-\text{inter1}} + L_{C-\text{inter2}}. \tag{8}$$

We can leverage the high-level semantic feature representations between two modalities for cross-modal alignment, and we achieve cross-modal heterogeneous data alignment by computing the minimization $L_{C-\text{inter}}$.

(2) Intramodal. To ensure the representativeness of the semantic information within the modality and to reduce semantic feature loss, we also perform intramodal constraints within the same modality, and we align the features extracted from the original image with the higher-level semantic representations encoded by the auto-encoder. Ensuring representability and completeness of high-level semantic information within a modality by minimizing $L_{C-\text{intra}}$, we construct the image modal feature matrix as $S_{F_{I,I}}$ after auto-encoder to obtain the features of the hidden state, which is denoted by $S_{\text{Enc}-I}$. The text features are also represented by $S_{F_{T,T}}$ for the original extracted features and $S_{\text{Enc}-T}$ for the features decoded by auto-encoder. We define the intramodal losses as follows:

$$L_{C-\text{intra}} = \sum \left\| S_{F^*} - S_{\text{Enc}-*} \right\|^2 * \in \{I, T\}. \tag{9}$$

Therefore, we construct a semantic alignment method with intramodal and intermodal alignment, which achieves intramodal semantic alignment by aligning the high-level semantic representation extracted by the visual coder and the text encoder with the compressed semantic features of the feature after auto-encode, ensuring that the high-latitude modal data can be restored with a small number of high-level features, aligning the heterogeneous data with the original modal features through the mapping of the decoder, enabling information interaction across the modal data, and achieving intramodal and intermodal alignment. We define the loss of cycle-alignment as follows:

$$L_C = L_{C-\text{inter}} + L_{C-\text{intra}}. \tag{10}$$

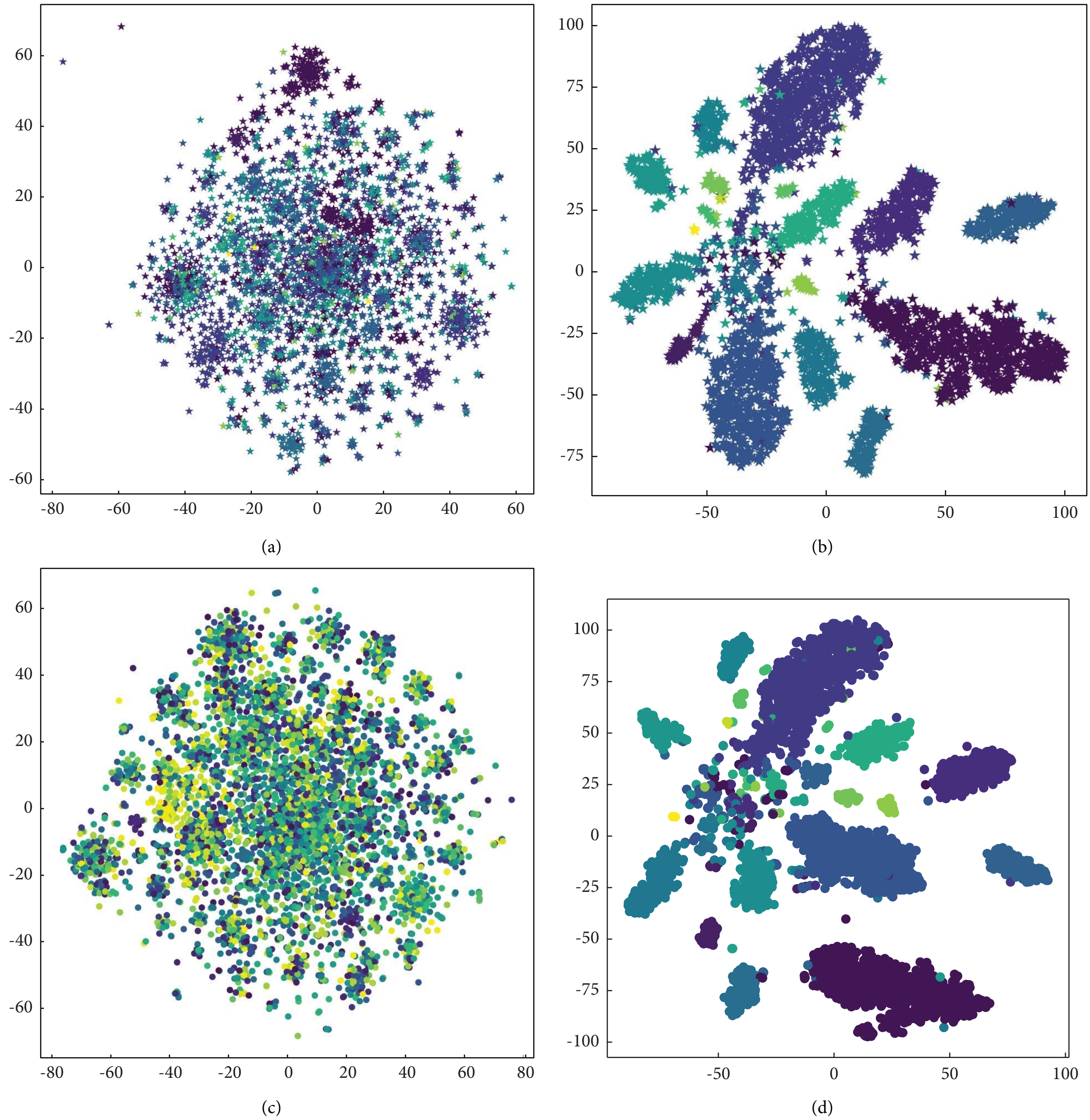
3.2.3. Hash Encoding Learning. After feature extraction and cycle semantic alignment, the semantic information of the text and visual data can be extracted and interlinked in a high-quality way. In the area of cross-modal retrieval, we aim to make semantically more similar heterogeneous data more closely related, by finding semantically related data samples from one modality in the dataset from query points in another modality according to a defined similarity metric. By converting the query points into a hash code, the corresponding modal information can be retrieved more quickly. With the AE (auto-encoder) mapping, we can fully extract the high-latitude feature encoding corresponding to each modality during the training phase. We perform the mapping of the hash encoding through the AE generated feature vector, and due to the feature extraction and reconstruction semantic operations, we use the true value to construct the hash encoding and generate hash codes via the tanh $(\cdot)$ function. We compute this pairwise cosine similarity matrix by defining them as S_{xy}^B, which is used to represent the generated hash matrix. The visualization of the feature generation hash encoding is shown in Figure 5. The hash matrix of text modalities is denoted as S_{TT}^B and the hash matrix of the image part is denoted as S_{II}^B. For the matrix elements, we calculate by using the following cosine similarity:

$$S_{xy}^B (i, j) = \cos\left(b_{x,i}, b_{y,j}\right). \tag{11}$$

In addition inspired by [22], to make fuller use of the semantic information described jointly by image text pairs, we construct cross-modal hash code similarity matrices where colinear image text pairs have the most similar labels or categories compared to other modal data, and the elements on the diagonal are better as they should be closer to 1, decomputed into hash codes for image text pairs, and minimizing the loss of colinear instances as follows:

$$\min_{B_I, B_T} \sum_{i=1}^{m} \left\| 1 - S_{I,T}^B \right\|^2. \tag{12}$$

Regarding other elements, we use diagonal similarity loss to bridge the connection between different modalities, e.g.,



(a)

(b)

(c)

(d)

FIGURE 5: t-SNE visualization of the data on the Flickr-25K. (a) Original image features. (b) Image encoded feature distribution. (c) Original text features. (d) Text encoded feature distribution. In the figure, the circle (○) and star (∗) denote the representation of text and image samples, respectively, and different colors denote the representation with different semantic categories.

the same pair of image text similarity should be independent of location information and only related to feature information, bridging the semantics of the image text pairs together by minimizing the diagonal loss, which we define as follows:

$$\min_{B_I,B_T} = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|S_{I,T}^{B}(i,j) - S_{I,T}^{B}(j,i)\right\|^2. \tag{13}$$

The total loss on $S_{I,T}^{B}$ is as follows:

$$\begin{aligned}\min_{B_I,B_T} L_S = {} & \sum_{i=1}^{m}\left\|1 - S_{I,T}^{B}(i,i)\right\|^2 \\ & + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|S_{I,T}^{B}(i,j) - S_{I,T}^{B}(j,i)\right\|^2.\end{aligned} \tag{14}$$

After autoencoder encoding, we map the obtained features to our hash codes through hash functions, and we use these hash codes to construct the similarity matrices $S_{I,I}^{B}$ and $S_{T,T}^{B}$. In addition, we introduce a new similarity matrix that we obtained by mapping the hash function $S^B_{I,T}$, which is constructed from image-text labels. We do not use labels for bootstrapping in the training phase, but introducing label information mainly to calculate the hash loss.

Whereas hashing methods can speed up the retrieval process, mapping truth-valued features to hash codes still results in some missing information, leading to suboptimal solutions for retrieval. In hash encoding learning, we also need to pay attention to the semantic relationships between data from different modalities, and similarity information across modalities is a central task in cross-modal retrieval. Based on this, we align the features within individual modalities with the generated hash codes to ensure that the generated hash codes are more realistic representations of the original data information. k is the modal adjustment parameter that allows more flexibility to ensure our semantic similarity.

$$L_{H-\text{intra}} = \sum \left\| k S^F_{x,x} - S^B_{x,x} \right\|^2 x \in \{I, T\}. \tag{15}$$

We are constructing a joint feature matrix that integrates the text feature matrix with the image feature matrix in a weighted way, which is represented by only one common matrix $S^F_{I,T}$. The ɑ is a hyperparameter that can be used to weight the feature matrix of images and text.

$$S^F_{I,T} = \alpha S^F_{I,I} + (1 - \alpha) S^F_{T,T}. \tag{16}$$

In optimising our hash encoding based on matrix alignment.

$$L_{H-\text{inter}} = \left\| k S^F_{I,I} - S^B_{I,I} \right\|^2 + \left\| k S^F_{I,T} - S^B_{T,T} \right\|^2 + \left\| k S^F_{I,T} - S^B_{I,T} \right\|^2. \tag{17}$$

The total loss between modes is as follows:

$$L_H = L_{H-\text{inter}} + \beta L_{H-\text{intra}}. \tag{18}$$

3.3. Optimisation.

We combine these losses to construct our total loss function as follows:

$$\min_{B_I, B_T} L = L_C + L_S + L_H. \tag{19}$$

Moreover, during our training process, the cyclic semantic interaction module uses truth codes, and during the training process, if the truth codes are converted into hash codes, some information will be lost, and the truth features are more conducive to the training of the model, and the truth codes generated after multiple modal interactions are closer to the hash codes. However, the generated truth codes cannot be gradient-derived because they are discrete values. To solve this problem, inspired by ($\lim_{\eta \longrightarrow \infty} \mathit{than}(\eta x) = \operatorname{sign}(x)$), we transform them into binary hash codes via $\tanh(.)$ with the following function definition:

$$B_I = \tanh\left(\eta H_I\right) B_T = \tanh\left(\eta H_T\right). \tag{20}$$

The proposed CCAH algorithm is shown in Algorithm 1.

4. Experiment

Datasets: our experiments were tested on three cross-modal retrieval datasets, including MIRFlickr-25K [32], NUS-WIDE [33], and MS COCO [34], to validate the effectiveness of our proposed model. The datasets are described as follows:

> MIRFlickr-25K: MIRFlickr contains 25,000 image-text pairs collected from the Flickr website. Each image text pair is saved as an instance. And for text patterns, after DJSRH [14], each text will be sorted and tagged with occurrence characteristics and transformed into a BOW (bag-of-words) vector.
>
> NUS-WIDE: NUS-WIDE consists of 269,648 pairs of multimodal data containing 81 categories, with each multimodal instance containing an image and corresponding label. For simple processing, we selected the 10 most frequent categories from the original 81 categories and the 186,577 tagged instances in all pairs. The text of each instance was represented as a 500-dimensional bag-of-words (BOW) vector. We collated the index vector of the most frequent 1,000 text labels.
>
> MS COCO: MS COCO was originally collected for the image understanding task and contains 123,287 images. For each image, a text description and a 91-dimensional semantic label are given. The experiment contains 87,081 images with category information and uses a 2,000-dimensional bag-of-words vector to represent the textual information. Of these, 5,000 image-text pairs were randomly selected as the query set and the remaining image-text pairs were used as the retrieval set. For the training set, 10,000 pairs were randomly sampled from the retrieval set.

4.1. Implementation Details.

We used CLIP as a feature extractor for image modality and GAT as a feature extractor for text modality. We used cyclic modal interaction to achieve semantic alignment within and between modalities (Intramodal and intermodal). We use hidden features of one modality to reconstruct features of another modality and carefully set some hyperparameters α, β, k to assist learning. We analyze the sensitivity of these parameters based on experiments. Finally, we selected our parameters as $\alpha = 0.8, \beta = 0.2$, and $k = 1.5$, batch-size is 16, the learning rate is 0.005 for both image and text modalities, the SGD optimization strategy is used, and the weight decay is set to 5×10^{-4}.

4.2. Baseline and Validation.

Evaluation criteria: we use three cross-modal common datasets, MIRFlickr-25K, NUS-WIDE, and MS COCO to validate our model. For MIRFlickr-25K and NUS-WIDE, we follow [14, 16, 17] and sample 2,000 instances as query points and the remainder as query database. Due to the overwhelming amount of data in MIRFlickr-25K and NUS-WIDE, we randomly sampled from one of the datasets in the database set for training. For fairness in training, we took some instances from each class

```
Require: Image set $I$; text set $T$;
        Batch size set $m$, hash code length $c$, Max epoch $E$.
Ensure: Deep Feature extract functions $F_T$, and $F_I$;
        encoder function set Enc – $(I/T)(*)$, and $* \in \{F_I, F_T\}$;
        Hash coding functions than$(\cdot)$, and $. \in \{B_T, B_I\}$.
Initialize the pretrained extractor parameters: $k, \alpha, \beta$.
While$e$ in $E$do
        $\eta = \sqrt{e}$;
        Extract the depth characteristics of each mode: $F_*$, $* \in \{I, T\}$;
        Encode the features to get the hidden states, by Enc$(*)$;
        Using the hidden states to generate truth matrix and hash codes;
        Decode the hidden states to generate heterogeneous features $F_I'$ and $F_T'$
        Calculate the objective function;
        Back propagate the gradient with the chain rule;
        Update the whole parameters;
end while
Return$F_I(.; \theta_I)$and $F_T(.; \theta_T)$
```

Algorithm 1: CLIP-based cycle alignment hashing for unsupervised vision-text retrieval.

in the first round of training and randomly sampled them in the remaining stages. In the MS COCO dataset, we take 10,000 instances as the retrieval set and the remainder makes up the database set. In our experiments, we take MAP and precision @ top-curves as the model judging criteria.

To validate our CCAH model, we compare it with some common cross-modal approaches. Shallow cross-modal Hashing: CVH [1], IMH [15], LCMH [35], CMFH [16], LSSH [9], RFDH [36], FSH [37], and STMH [38]. Deep cross-modal Hashing: DBRC [24], UDCMH [17], DJSRH [14], DSAH [22], JDSH [39], MGAH [40], JIMRH [23], HNH [25], and DUCH [41]. The results of CCAH compared to other models are shown in Figure 6.

We compare with previous work on the MIRFlickr-25K and NUS-WIDE datasets, where we used a benchmark of MAP@50. The total retrieval accuracy of our CCAH model demonstrates better results than previous work in different coding lengths as shown in Table 1.

As can be seen, our experimental data demonstrate excellent results on two widely used datasets, with significant gains in both image retrieval text and text retrieval image on MIRFlickr-25K, and slightly worse results for image retrieval text on the NUS-WIDE dataset, but significant gains in text retrieval image accuracy, and gains in overall retrieval accuracy. The NUS-WIDE (tc-10) dataset was used, taking the most common 10 classes as the composition of the dataset. As the NUS-WIDE dataset is relatively large, it is not possible to ensure that the classes of the sample points taken are equal when sampling the sample points, and the data is more sparse when constructing the adjacency matrix, leading to a reduction in the efficiency of image retrieval of the text. To validate our theory, guided by DAEH [42], we tested again on the MS COCO dataset, which uses class 81. We used MAP@5000 to evaluate our model and the results are shown in Table 2.

4.3. Ablation Experiment. We experimentally validate the effect of different modules on the accuracy and we validate the model on the MIRFlickr dataset for 128 bits. We have also made other attempts. In the encoding and compression phase, we adopt a two-way model where the compressed vector reconstructs both its original features and the original features of the heterogeneous data, rather than just the features of the heterogeneous data. We validated this on the MIRFlickr and NUS-WIDE datasets. The results show that if we add homogeneous feature reconstruction, there is a relative 1% improvement in image retrieval of text, but the accuracy of text retrieval of images decreases (Table 3).

In Table 4, we perform ablation experiments on different modules to demonstrate the effectiveness of our proposed method.

4.4. Visualization of the Learned Representation. To visualize the effectiveness of the proposed CCAH, we use t-SNE to visualize the learned representation of images, text on the Flickr-25K dataset (Figure 5). The original feature representation of the images and text are shown in Figures 5(a) and 5(c), respectively. It can be seen that the distributions of these modalities have large differences and it is difficult to distinguish the samples by the original representations. Figures 5(b) and 5(d) gives the distribution of the learned representations of the images and text, respectively. It can be seen from the figures that the proposed CCAH method helps to distinguish samples with different semantic classes and some clusters show distinguished intervals.

4.5. Hyperparameter Sensitivity. We further validated our parameters k, α, and β on three datasets using 128 bits coding lengths. k is the influence factor by which we optimize our hash matrix with the eigenvalues into an alignment of the hash code, and we find that the best results are obtained when $k = 1.5$. α is the parameter for aligning images and text across modalities. It is known that image modalities contain richer semantic features than text modalities (Figure 1), so when weighting images and text, the image component is weighted more than the text, and our model

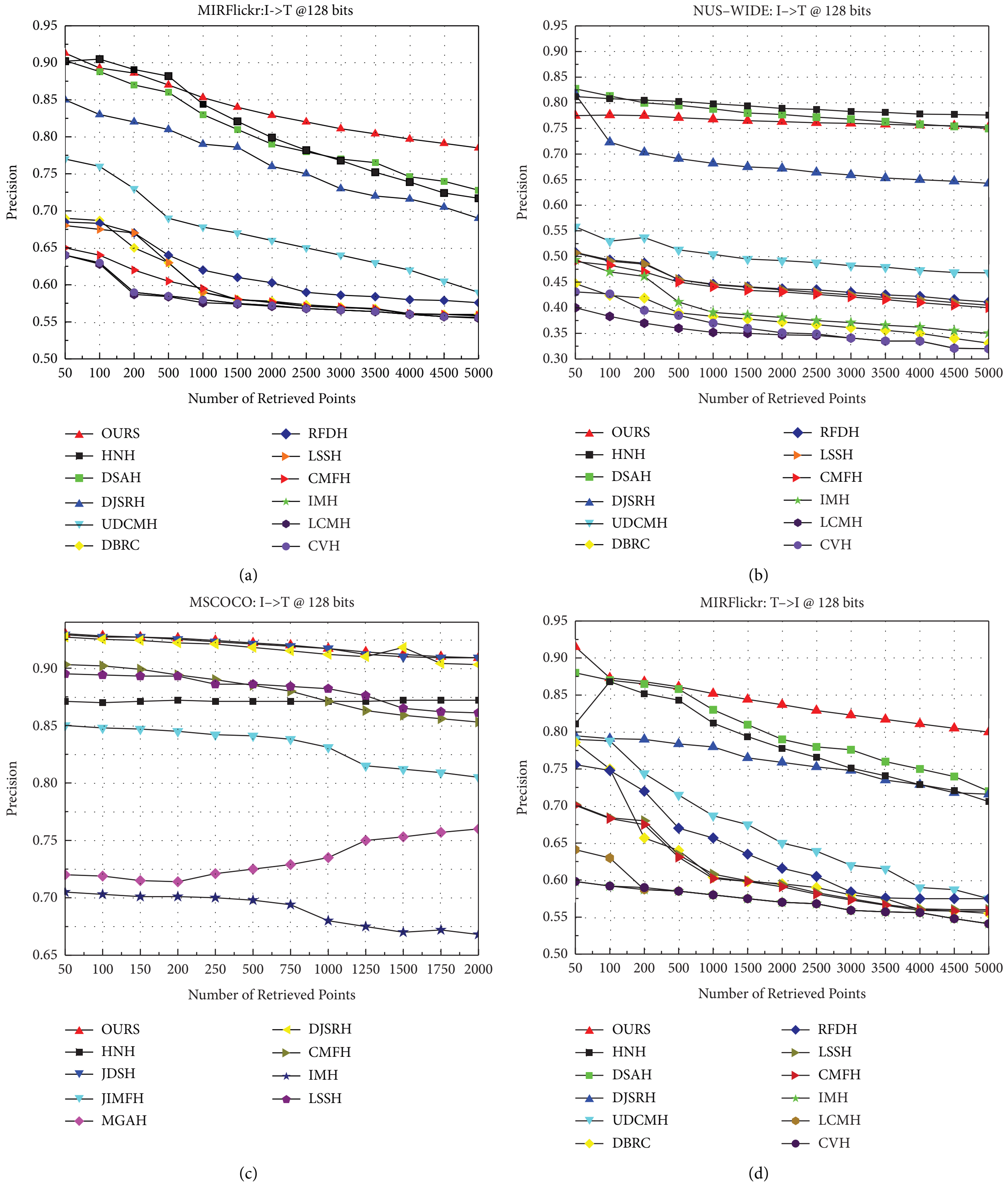
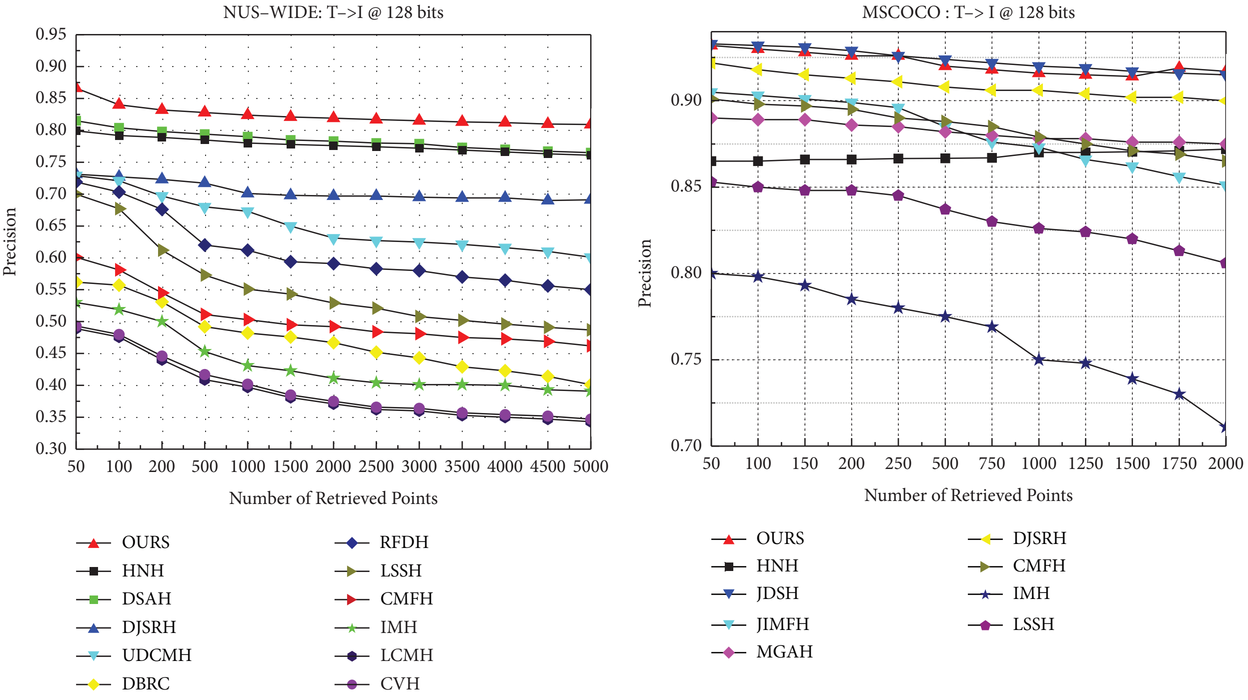


Figure 6: Continued.



(e) (f)

Figure 6: MAP@topK curves on MIRFlickr-25K, NUS-WIDE, and MS COCO.

Table 1: Comparison results on mean accuracy (MAP@50) for different code lengths under the Flickr-25K and NUS-WIDE dataset.

Dataset	Flickr-25K								NUS-WIDE							
Task	I->T				T->I				I->T				T->I			
Method	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CVH	0.606	0.599	0.596	0.598	0.591	0.583	0.576	0.576	0.372	0.362	0.406	0.39	0.401	0.384	0.442	0.432
IMH	0.612	0.601	0.592	0.579	0.603	0.595	0.589	0.58	0.47	0.473	0.476	0.459	0.478	0.483	0.472	0.462
LCMH	0.559	0.569	0.585	0.593	0.561	0.569	0.582	0.582	0.354	0.361	0.389	0.383	0.376	0.387	0.408	0.419
CMFH	0.621	0.624	0.625	0.627	0.642	0.662	0.676	0.685	0.455	0.459	0.465	0.467	0.529	0.577	0.614	0.645
LSSH	0.584	0.599	0.602	0.614	0.618	0.626	0.626	0.628	0.481	0.489	0.507	0.507	0.455	0.459	0.468	0.473
RFDH	0.632	0.636	0.641	0.652	0.681	0.693	0.698	0.702	0.488	0.492	0.494	0.508	0.612	0.641	0.658	0.68
DBRC	0.617	0.619	0.62	0.621	0.618	0.626	0.626	0.628	0.424	0.459	0.447	0.447	0.455	0.459	0.468	0.473
UDCMH	0.689	0.698	0.714	0.717	0.692	0.704	0.718	0.733	0.511	0.519	0.524	0.558	0.637	0.653	0.695	0.716
DJSRH	0.810	0.843	0.862	0.876	0.786	0.822	0.835	0.847	0.724	0.773	0.798	0.817	0.712	0.744	0.771	0.789
DSAH	0.863	0.877	0.895	0.903	0.846	0.860	0.881	0.882	**0.775**	**0.805**	**0.818**	**0.827**	0.770	0.790	0.804	0.815
HNH	0.853	**0.883**	0.895	0.902	0.833	0.854	0.868	0.878	0.582	0.747	0.800	0.816	0.423	0.743	0.781	0.780
CCAH	**0.863**	0.879	**0.899**	**0.910**	**0.891**	**0.908**	**0.914**	**0.913**	0.715	0.754	0.775	0.787	**0.834**	**0.851**	**0.864**	**0.874**

achieves the best results when $\alpha = 0.8$. β is the parameter that balances the hash encoding with the original features and also boosts the intramodal and intermodal coefficients. The visualization of hyperparametric sensitivity is shown in Figure 7.

4.6. Comparing Other Models. On the 3 cross-modal common datasets mentioned above, our results are significantly improved compared to other models, and our total retrieval accuracy in top-k exceeds previous methods in all cases. We added the GAT network, which successfully constructs adjacency matrices employing graph neighbors to attentionally boost semantic feature-poor text modalities with higher accuracy compared to traditional bag-of-words features. Using CLIP to extract image features, the CLIP large-scale pretrained model can extract features from images at a finer level. We construct a cyclic semantic alignment module to construct the semantic features of the heterogeneous modes by using the hidden state vector of each mode from the self-encoder, compared to using a binary code to construct the features, the true value information is more representative of the mode features and a lot of useful information is lost by using the binary code.

Table 2: Comparison results on mean accuracy (MAP@5000) for different code lengths under the Flickr-25K, NUS-WIDE, and MSCOCO dataset.

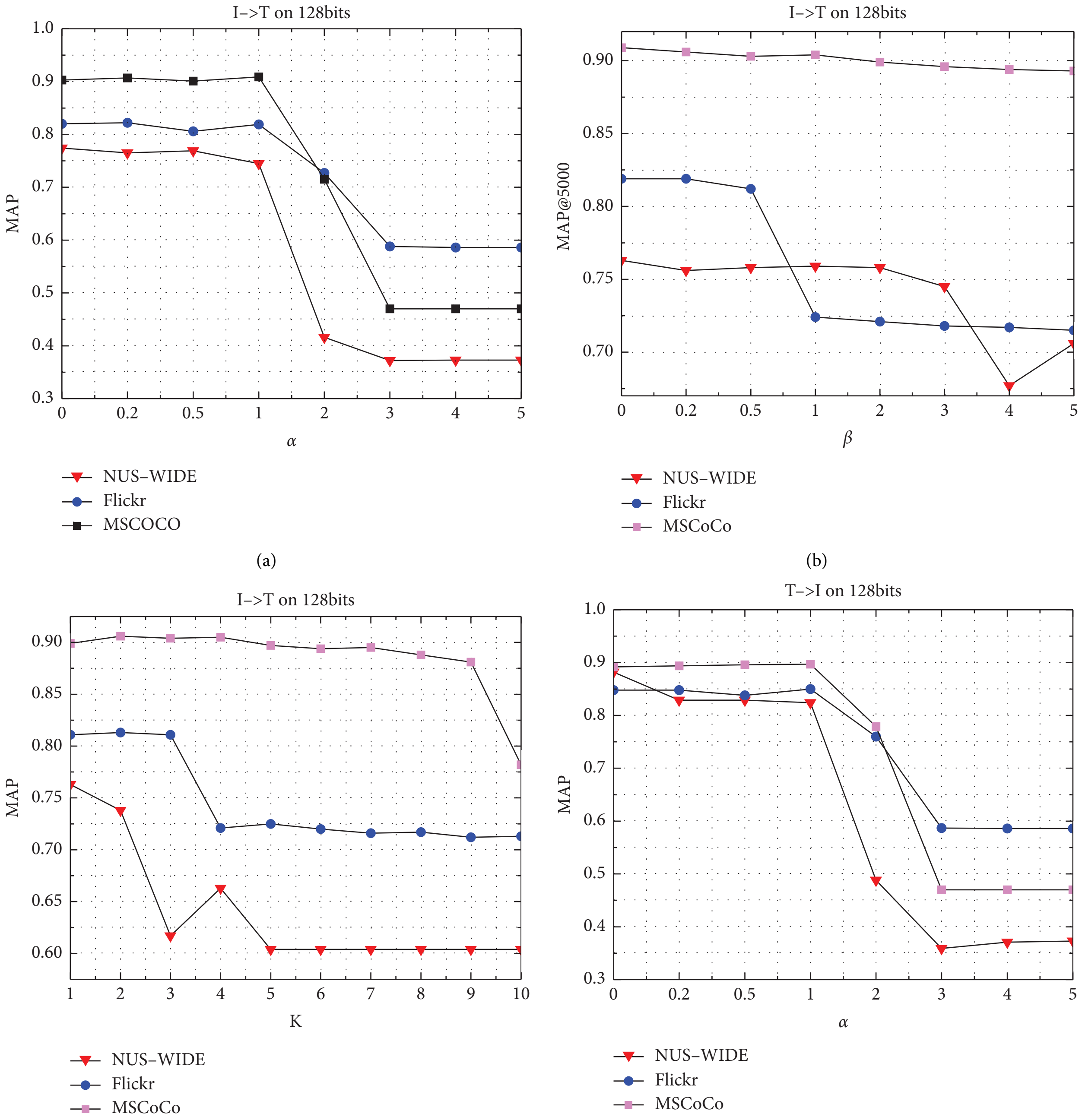
Task		MIRFLickr-25K				NUS-WIDE				MSCOCO			
	Method	16	32	64	128	16	32	64	128	16	32	64	128
I->T	IMH	0.681	0.659	0.643	0.633	0.607	0.623	0.619	0.591	0.737	0.687	0.681	0.659
	LSSH	0.675	0.677	0.682	0.684	0.678	0.706	0.703	0.694	0.813	0.832	0.838	0.848
	STMH	0.566	0.585	0.619	0.659	0.409	0.429	0.468	0.482	0.646	0.687	0.653	0.738
	CMFH	0.686	0.692	0.701	0.718	0.635	0.664	0.699	0.731	0.725	0.757	0.777	0.816
	FSH	0.659	0.678	0.684	0.706	0.578	0.596	0.631	0.634	0.748	0.770	0.794	0.810
	RFDH	0.636	0.648	0.658	0.681	0.551	0.572	0.608	0.649	0.690	0.710	0.749	0.782
	DJSRH	0.673	0.701	0.730	0.744	0.587	0.671	0.717	0.743	0.754	0.815	0.861	0.870
	MGAH	0.631	0.649	0.658	0.692	0.601	0.677	0.715	0.586	0.780	0.807	0.814	0.752
	JIMRH	0.611	0.622	0.633	0.632	0.493	0.516	0.551	0.588	0.660	0.706	0.732	0.756
	JDSH	0.725	0.731	0.752	0.761	0.678	0.724	0.743	0.756	0.690	0.758	0.888	0.890
	DSAH	0.639	0.766	0.779	0.789	**0.724**	**0.753**	**0.772**	**0.778**	0.850	0.881	0.900	0.900
	HNH	0.730	0.745	0.738	0.721	0.684	0.721	0.740	0.737	0.830	0.855	0.868	0.850
	DUCH	0.667	0.688	0.706	0.723	0.686	0.714	0.728	0.747	0.847	0.866	0.876	0.883
	CCAH	**0.783**	**0.801**	**0.815**	**0.815**	0.715	0.733	0.753	0.764	**0.874**	**0.898**	**0.907**	**0.907**
T->I	IMH	0.681	0.667	0.654	0.640	0.626	0.644	0.638	0.617	0.768	0.717	0.715	0.694
	LSSH	0.648	0.653	0.662	0.660	0.567	0.587	0.624	0.628	0.708	0.745	0.779	0.800
	STMH	0.643	0.674	0.690	0.694	0.581	0.611	0.645	0.675	0.686	0.769	0.811	0.833
	CMFH	0.661	0.669	0.679	0.695	0.609	0.641	0.672	0.696	0.757	0.789	0.809	0.838
	FSH	0.682	0.697	0.702	0.725	0.609	0.649	0.665	0.668	0.769	0.791	0.809	0.826
	RFDH	0.625	0.646	0.654	0.663	0.551	0.568	0.592	0.630	0.701	0.717	0.741	0.777
	DJSRH	0.675	0.691	0.698	0.712	0.601	0.656	0.707	0.719	0.759	0.832	0.862	0.869
	MGAH	0.627	0.648	0.625	0.632	0.590	0.613	0.645	0.688	0.747	0.772	0.768	0.845
	JIMRH	0.647	0.647	0.657	0.659	0.584	0.586	0.610	0.626	0.728	0.767	0.779	0.802
	JDSH	0.699	0.719	0.724	0.735	0.674	0.715	0.711	0.718	0.758	0.829	0.895	0.895
	DSAH	0.646	0.754	0.759	0.758	0.668	0.716	0.748	0.745	0.854	0.886	0.890	0.891
	HNH	0.723	0.720	0.706	0.700	0.671	0.699	0.696	0.693	0.839	0.863	0.867	0.851
	DUCH	0.652	0.668	0.681	0.697	0.662	0.694	0.709	0.713	0.860	0.885	**0.898**	**0.903**
	CCAH	**0.803**	**0.826**	**0.835**	**0.848**	**0.764**	**0.798**	**0.806**	**0.826**	**0.870**	**0.886**	0.890	0.897

Table 3: Reconstructive experimental ablation analysis.

Bits	MIRFlickr-25K				NUS-WIDE			
	16	32	64	128	16	32	64	128
$I \longrightarrow T$	0.869	0.898	0.913	0.919	0.743	0.755	0.770	0.789
$T \longrightarrow I$	0.878	0.893	0.899	0.902	0.797	0.827	0.835	0.842

Table 4: MAP@50 results at MIRFlickr-25K and NUS-WIDE for ablation analysis.

Method	Firlickr-25K								NUS-WIDE							
	I->T				T->I				I->T				T->I			
Bits	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
CCAH	0.863	0.879	0.899	0.91	0.891	0.908	0.914	0.913	0.715	0.754	0.775	0.787	0.834	0.851	0.864	0.874
GAT	0.892	0.922	0.935	0.905	0.886	0.886	0.897	0.896	0.796	0.829	0.841	0.85	0.774	0.792	0.8	0.808
CLIP	0.869	0.871	0.888	0.901	0.89	0.895	0.906	0.908	0.735	0.759	0.781	0.793	0.836	0.842	0.858	0.869
CA	0.859	0.876	0.891	0.907	0.883	0.899	0.906	0.904	0.713	0.748	0.771	0.784	0.828	0.848	0.861	0.87
ALL	0.863	0.877	0.895	0.903	0.846	0.86	0.881	0.882	0.775	0.805	0.818	0.827	0.772	0.791	0.804	0.815



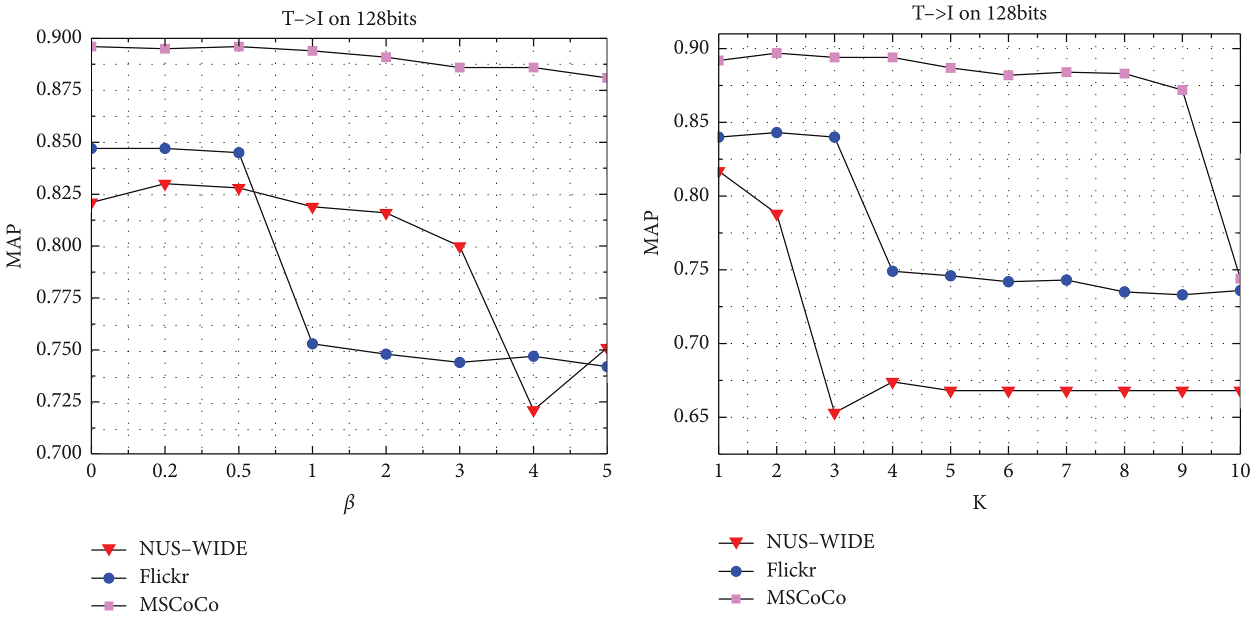
(a)

(b)

(c)

(d)

Figure 7: Continued.



(e)

(f)

Figure 7: Parametric sensitivity analysis on MIRFlickr, MS COCO, and NUS-WIDE datasets at 128 bits.

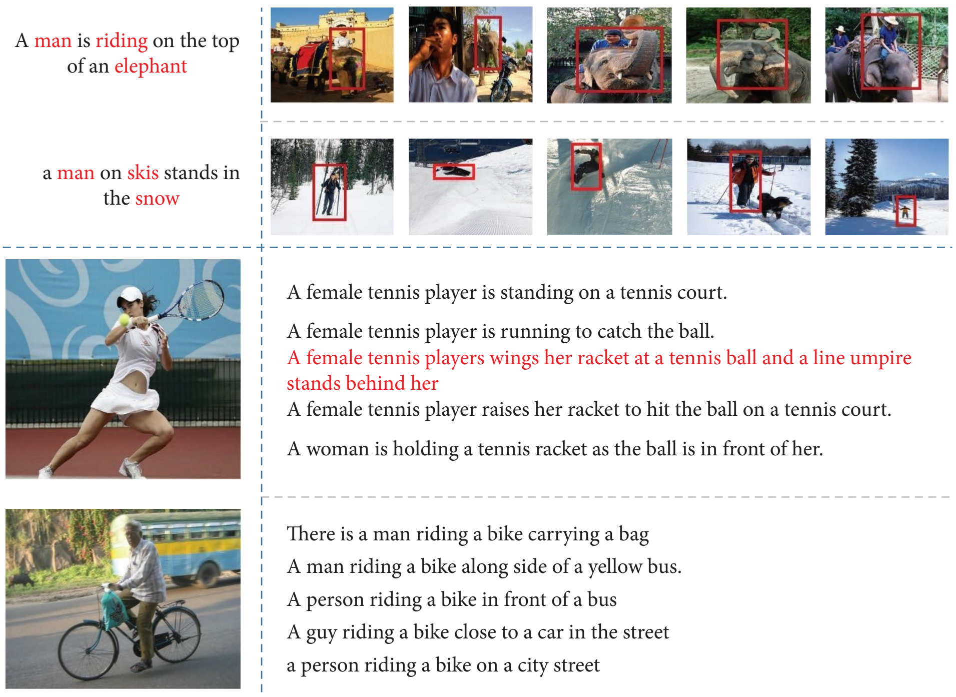

A man is riding on the top of an elephant
a man on skis stands in the snow
A female tennis player is standing on a tennis court.
A female tennis player is running to catch the ball.
A female tennis players wings her racket at a tennis ball and a line umpire stands behind her
A female tennis player raises her racket to hit the ball on a tennis court.
A woman is holding a tennis racket as the ball is in front of her.
There is a man riding a bike carrying a bag
A man riding a bike along side of a yellow bus.
A person riding a bike in front of a bus
A guy riding a bike close to a car in the street
a person riding a bike on a city street

Figure 8: Text query image and image retrieval text on the MS COCO dataset. The building boxes of images are manually labelled for readability. The image retrieval text is shown in red for incorrect retrieval results.

We perform validation of our model on the MS COCO dataset and we mark the detected images with manual regions. In text retrieved images, text marked in red is the feature word of the text (corresponding to the image marker region); in image retrieved text, text marked in red indicates that the retrieval result does not quite match the description of the image Figure 8.

5. Conclusion

In this paper, we propose a novel deep unsupervised cross-modal hashing method, CLIP-based cycle alignment hashing (CCAH) for unsupervised vision-text retrieval. We construct a cycle alignment module that allows for more flexible exploitation of high-level semantic information within and across modalities. To further bridge the gap between the two modalities, we use the hidden state vector of one modality to reconstruct the features of the other modality, enabling cross-modal data to be mutually characterized. Extensive experiments on three benchmark datasets show that CCAH outperforms several state-of-the-art methods in multimodal data retrieval tasks.

Data Availability

The data and code that support the results of this study are openly available in CCAH at https://github.com/CQYIO/CCAH.git.

Disclosure

A preprint has previously been released [43].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Mingyong Li proposed the model idea and guided the writing of the paper and participated in the revision of the paper. Longfei Ma is responsible for the paper and experimental implementation. Yewen Li is responsible for model result validation. Mingyuan Ge is responsible for data validation.

Acknowledgments

This work was partially supported by Chongqing Natural Science Foundation of China (Grant nos. CSTB2022NSCQ-MSX1417), the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant no. KJZD-K202200513), and Chongqing Normal University Fund (Grant no. 22XLB003).

References

[1] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Washington, DC, USA, July 2011.

[2] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 769–790, 2018.

[3] Z. Lin, G. Ding, M. Hu, and W. Jianmin, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3864–3872, Boston, MA, USA, June 2015.

[4] D. Wu, Z. Lin, B. Li, and M. Ye, "Deep supervised hashing for multi-label and large-scale image retrieval," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 150–158, New York, NY, USA, June 2017.

[5] D. Wu, J. Liu, B. Li, and W. Wang, "Deep index-compatible hashing for fast image retrieval," in *Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, San Diego, CA, USA, July 2018.

[6] E. Yang, C. Deng, W. Liu, and X. Liu, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, February 2017.

[7] M. M. Bronstein, A. M. Bronstein, F. Michel, and P. Nikos, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3594–3601, IEEE, San Francisco, CA, USA, June 2010.

[8] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2469–2479, 2016.

[9] B. Wu, Q. Yang, W. S. Zheng, Y. Wang, and J. Wang, "Quantized correlation hashing for fast cross-modal search," in *Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence*, Washington, DC, USA, July 2015.

[10] W. Gu, X. Gu, J. Gu, and B. Li, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 159–167, New York, NY, USA, June 2019.

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

[12] Z. Ye and Y. Peng, "Multi-scale correlation for sequential cross-modal hashing learning," in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 852–860, New York, NY, USA, October 2018.

[13] D. Zhang and W. J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, July 2014.

[14] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3027–3035, Seoul, Korea, October 2019.

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

[16] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2075–2082, Columbus, OH, USA, June 2014.

[17] G. Wu, Z. Lin, J. Han, and L. Liu, "Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval," in *Proceedings of the International Joint Conferences on Artifical Intelligence*, pp. 2854–2860, Vienna, Austria, July 2018.

[18] C. Li, C. Deng, L. Wang, and D. Xie, "Coupled cyclegan: unsupervised hashing network for cross-modal retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 176–183, Washington, DC, USA, March 2019.

[19] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[20] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018.

[21] L. Jin, X. Shu, K. Li, Z. Li, G. J. Qi, and J. Tang, "Deep ordinal hashing with spatial attention," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2173–2186, 2019.

[22] D. Yang, D. Wu, W. Zhang, and B. Li, "Deep semantic-alignment hashing for unsupervised cross-modal retrieval," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pp. 44–52, New York, NY, USA, June 2020.

[23] D. Wang, Q. Wang, L. He, X. Gao, and Y. Tian, "Joint and individual matrix factorization hashing for large-scale cross-modal retrieval," *Pattern Recognition*, vol. 107, Article ID 107479, 2020.

[24] D. Hu, F. Nie, and X. Li, "Deep binary reconstruction for cross-modal hashing," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 973–985, 2019.

[25] P. F. Zhang, Y. Luo, Z. Huang, X. S. Xu, and J. Song, "High-order nonlocal hashing for unsupervised cross-modal retrieval," *World Wide Web*, vol. 24, no. 2, pp. 563–583, 2021.

[26] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[27] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[28] W. Kim, B. Son, and I. Kim, "Vilt: vision-and-language transformer without convolution or region supervision," in *Proceedings of the International Conference on Machine Learning, PMLR*, pp. 5583–5594, New York, NY, USA, February 2021.

[29] A. Radford, J. W. Kim, C. Hallacy, R. Aditya, and G. Gabriel, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning, PMLR*, pp. 8748–8763, New York, NY, USA, February 2021.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and S. Gelly, "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[31] P. Veličković, G. Cucurull, A. Casanova, and Y. Bengio, "Graph attention networks," 2017, https://arxiv.org/abs/1710.10903.

[32] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, New York, NY, USA, October 2008.

[33] T. S. Chua, J. Tang, R. Hong, and Y. Zheng, "Nus-wide: a real-world web image database from national university of Singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 1–9, New York, NY, USA, July 2009.

[34] T. Y. Lin, M. Maire, S. Belongie, and H. James, "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, Springer, New York, NY, USA, February 2014.

[35] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 143–152, New York, NY, USA, October 2013.

[36] D. Wang, X. Gao, X. Wang, L. He, and B. Yuan, "Multimodal discriminative binary embedding for large-scale cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4540–4554, 2016.

[37] H. Liu, R. Ji, Y. Wu, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7380–7388, Honolulu, HI, USA, July 2017.

[38] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Washington, DC, USA, July 2015.

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

[40] J. Zhang and Y. Peng, "Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 174–187, 2020.

[41] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing," 2022, https://arxiv.org/abs/2201.08125.

[42] Y. Shi, Y. Zhao, X. Liu et al., "Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7255–7268, 2022.

[43] L. Ma, Y. Li, and M. Ge, "Clip-based cycle alignment hashing for unsupervised vision-text retrieval," *Research Square*, 2022.