applied sciences

MDPI

Article

TDCMR: Triplet-Based Deep Cross-Modal Retrieval for Geo-Multimedia Data

Jiagang Song [1,†], Yunwu Lin [2,*], Jiayu Song [2,*], Weiren Yu [3,4] and Leyuan Zhang [1,5]

[1] School of Computer Science and Engineering, Guangxi Normal University, Guiling 541004, China; songjg@stu.gxnu.edu.cn (J.S.); seaton@stu.gxnu.edu.cn (L.Z.)
[2] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[3] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China; ywr0708@hotmail.com
[4] Department of Computer Science, University of Warwick, Warwick CV4 8UW, UK
[5] School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia
* Correspondence: lywcsu@csu.edu.cn (Y.L.); jiayusong@csu.edu.cn (J.S.); Tel.: +86-185-7491-9704 (Y.L.)
† Current address: College of Computer Science and Electronic Engineering, Guangxi Normal University, Guiling 541004, China.

Abstract: Mass multimedia data with geographical information (geo-multimedia) are collected and stored on the Internet due to the wide application of location-based services (LBS). How to find the high-level semantic relationship between geo-multimedia data and construct efficient index is crucial for large-scale geo-multimedia retrieval. To combat this challenge, the paper proposes a deep cross-modal hashing framework for geo-multimedia retrieval, termed as Triplet-based Deep Cross-Modal Retrieval (TDCMR), which utilizes deep neural network and an enhanced triplet constraint to capture high-level semantics. Besides, a novel hybrid index, called TH-Quadtree, is developed by combining cross-modal binary hash codes and quadtree to support high-performance search. Extensive experiments are conducted on three common used benchmarks, and the results show the superior performance of the proposed method.

Keywords: geo-multimedia; nearest neighbor query; cross-modal hashing; triplet loss; TH-Quadtree

Citation: Song, J.; Lin, Y.; Song, J.; Yu, W.; Zhang, L. TDCMR: Triplet-Based Deep Cross-Modal Retrieval for Geo-Multimedia Data. *Appl. Sci.* **2021**, *11*, 10803. https://doi.org/10.3390/app112210803

Academic Editor: Adriano Ribolini

Received: 10 September 2021
Accepted: 11 November 2021
Published: 16 November 2021

1. Introduction

With the rapid development of mobile internet, social networks, and Location-Based Service (LBS), large numbers of multimedia data [1] with geographical information (a.k.a geo-multimedia) [2], such as text, image [3,4], and video [5–8], are collected and stored on the internet. As an important data resource, geo-multimedia data is used to support for location-based recommendation, accurate advertising and data search. Nearest neighbor spatial keyword query (NNSKQ) is a very important retrieval technique in LBS applications, which only focuses on location information and keyword information to find spatial objects. That means it is limited to structured data or text modality [9–11], which cannot be directly applied to geo-multimedia data [12–14]. However, the traditional multi-modal retrieval techniques ignore the geographic location information. To solve this dilemma, many researchers have tried to integrate multi-modal information into the query and proposed an effective nearest neighbor query method for geo-multimedia data [15].

In addition, two groups of tasks, i.e., cross-modal retrieval and spatial textual query, interlock with geo-multimedia retrieval. Cross-modal retrieval [16] is a hotspot in the multimedia community, which is aiming to search multimedia instance by queries of different modalities [17–20]. The challenge of cross-modal retrieval is to diminish semantic gap between different modalities, which is the main obstacle to measure cross-modal semantic similarity. Recently, lots of deep learning-based works are proposed [13,14,21–24], which outperform the traditional hand-crafted feature-based approaches. On the other

hand, in order to better apply on large-scale multimedia database, many researchers focus on cross-modal hashing method to reduce the search and storage cost [25–28]. Compact binary hash codes are generated from multimedia instances by deep neural networks, which contains semantic information. The other task, spatial textual query, is to do a keyword-based query by the aid of geographical information to reduce the candidate set substantially. Several important researches [29–33] are proposed in the last decade, which are support efficient LBS. Recently, some studies [2,34,35] extended spatial indexes to multimedia data, such as geo-images top-k query, k nearest neighbor query, spatial visual similarity join, and geo-image reverse query. These works make full use of geo-multimedia data via hybrid indexes to organize geo-multimedia instances through both contents and geo-locations.

Motivation. Although great progress has been made, there are still two challenges in cross-modal retrieval for geo-multimedia data. One challenge is insufficient semantic similarity learning of geo-multimedia data lead to inaccuracy of retrieval. Two problems of cross-modal hashing [36–38] cannot be neglected: On the one hand is how to extract the multi-modal features effectively to capture high-level semantic features [39]. Traditional methods use hand-crafted features for hash function learning, which cannot represent high-level semantics efficiently. Compared with them, the deep neural network can be used to extract the high-level semantics. On the other hand, how to establish the semantic relationships effectively to narrow the semantic gap. The existing cross-modal hashing methods use pairwise similarity constraints (pairwise label) as supervision information, so that the distance between similar image-text pairs is less than that of dissimilar image-text pairs [40]. That means the relative semantic relationship between cross-modal data is lost, which limits the semantic representation capability during hashing learning. The other challenge is *inefficient index and retrieval algorithm in massive geo-multimedia database*. To overcome this problem, in Reference [41], a novel hybrid index called GMR-Tree is developed that is an extension of R-Tree by integrating cross-modal representation. However, this work ignores the cross-modal hashing representation that can enhance the search efficiency significantly. The GMR-Tree-based search algorithm cannot be directly used for cross-modal hashing retrieval. Therefore, to address these two challenges, we try to improve the semantic similarity constraints to guide the deep neural networks and designed a hybrid geo-multimedia index to organize cross-modal hash codes. Last but not least, we introduce one detailed example to describe this problem more vividly as follows.

Example 1: As illustrated in Figure 1, a user wants to buy a blue shoe from a nearby shop. Unfortunately, he does not know the brand and of the shoe, and it is hard for him to use some words to describe the features of this shoe. Obviously, it is hard for him to find suitable shoes from nearby shops with limited information. In this case, he can input an image, which describes this blue shoe, and his location information by his mobile as a spatial multimedia k nearest neighbor query to a geo-tagged multimedia data retrieval system. According to his requirement, this system will return a result set, which contains k geo-multimedia data meeting his requirements. This result set indicates the user which shops have this kind of shoe and are close to his location.

Our Method. To this end, this paper proposes a novel efficient cross-modal hashing approach, termed as **T**riplet-based **D**eep **C**ross-**M**odal **R**etrieval (**TDCMR**). Specifically, a two-branch deep neural network-based backbone is integrated in the TDCMR framework, which is used to learn abstract semantic concepts. Besides, an improved triplet distance constraint is designed to capture multiple high-level similarities to capture semantic relationship among heterogeneous multi-modal data. Thus, the integrating of deep representation learning with an enhanced triplet distance constraint improves the cross-modal semantic learning performance. In addition, to realize efficient search on large-scale geo-multimedia data, a novel index, termed as TDCMR-Quadtree is proposed. It is a geo-semantic hybrid index integrated quadtree with cross-modal hash codes, which utilizes both geographic information and semantic similarity to realize candidate set prun-

ing. Based on the index, an efficient cross-modal nearest neighbor query algorithm is developed for geo-multimedia retrieval.

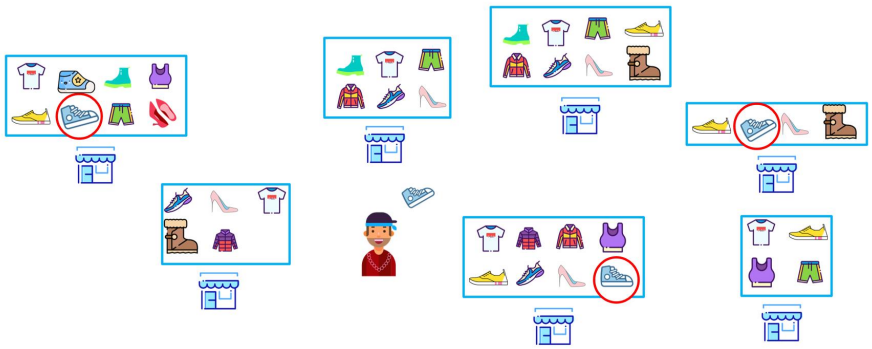


Figure 1. An example of a spatial multimedia k nearest neighbor query on a geo-tagged multimedia data retrieval system.

Contributions. The main contributions of our work are summarized as follows:

- We propose a triplet-based deep cross-modal hashing framework, named Triplet-based Deep Cross-Modal Retrieval (TDCMR), which aims to extract deep sample features to alleviate the semantic gap through a triplet deep neural network unified feature learning and hash learning process.
- We carefully apply the efficiency TDCMR algorithm to utilize the Quadtree to improve and enhance the search performance significantly, named TDCMR-Quadtree, which is a novel index framework and can improve the retrieval efficiency.
- We have conducted extensive experiments on three common used benchmarks, and the results demonstrate that our proposed method achieves very high performance.

Roadmap. In the remainder of this paper, we review the previous researches in Section 2. In Section 3, we introduce the definition of geo-multimedia data k nearest neighbor query and the related notions. In Section 4, we introduce the Triplet-based Deep Cross-Modal Retrieval (TDCMR) framework and its implementation. In Section 5, we evaluate our method on two geo-multimedia datasets. Finally, we conclude this paper in Section 6.

2. Related Work

2.1. Cross-Modal Hashing

Cross-Modal Hashing aims to map high-dimension data of different modalities into a common hash code space, in which the heterogeneous data realizes semantic representation and similarity measurement. With the rapid development of deep learning [21,22], deep learning-based cross-modal hashing [25–28] has made significant progress recently. It can effectively capture high-level information and explore semantic relevance to bridge modality gap [42]. To preserve the cross-modal similarities, a negative log-likelihood loss function is used by Deep cross-modal hashing (DCMH) [43], which performs feature learning and hash-code learning in an end-to-end learning framework. Pairwise Relationship Guided Deep Hashing (PRDH) [40] integrates different types of pairwise constraints to guide the hash code learning from intra-modality and inter-modality, respectively, and introduces additional decorrelation constraints to enhance the discriminative ability of each hash bit. To enhance the retrieval accuracy, Self-Supervised Adversarial Hashing (SSAH) [44] enhances the retrieval accuracy by jointly utilizing two adversarial networks, but its training cost is a little high, and its practical value is too low.

2.2. Spatial Related Data Retrieval

With the proliferation of local services and GPS-enabled mobile phones, there is a rapidly growing amount of spatio-textual data and increased need for spatial data retrieval,

so Spatial Keyword k Nearest Neighbor Query (sKkNN) is becoming an important type of query. Zhang et al. [29] designed a new index named IL-Quadtree and proposed an efficient algorithm to improve the performance of query. Cong et al. [30] proposed a new indexing framework, which adopts an invert file to search text data and employs R-tree to index closed objects, based on top-k aggregation problem. In order to further improve retrieval performance, Rocha-Junior et al. [32] invented spatial inverted index technology.

2.3. Quadtree

Quadtree is a tree data structure proposed by Raphael Finkel et al. in 1974. Spatial indices store and manage the spatial data sequentially according to the geographic location, shape and spatial relationship of spatial objects, which contain the identifier, pointer, and other description information of spatial objects [29]. Compared to conventional index types, spatial indices can efficiently handle spatial queries [45], such as how far two points differ, or whether points fall within a spatial area of interest. We use the quadtree index to organize the spatial data in this paper because of its simple structure and high retrieval efficiency. Quadtree is an extension of binary tree in high-dimensional space and has many variants, such as region quadtree, point quadtree, PR quadtree, MX quadtree, etc. The basic idea of the quadtree is to recursively subdivide a two-dimensional space into different levels of tree. Specially, it partitions the geographic space into four quadrants or regions in two orthogonal directions, and it recursively subdivide the subspaces until the tree reaches the maximum depth or the number of objects in the leaf node is less than or equal to the predetermined amount.

3. Preliminaries

In this section, we propose the formal definitions of geo-multimedia data similarity and k nearest neighbor query. Then, we review Siamese network and triple network, which are the basis of our work. Table 1 summarizes the frequently used notations in this paper.

Table 1. The summary of notations.

Notation	Definition
O	Space multimedia datasets
o	geo-multimedia Data Objects
$o.uID$	Identification code o spatial objects
$o.loc$	Geographical location information o spatial objects, composed of longitude $o.lng$ and latitude $o.lat$.
$o.m$	Multimedia Data Information o Space Objects
$o.v^T$	Cross-modular hash code for spatial objects o text modes
$o.v^I$	Cross-modular hash code for spatial objects o image modes
q	geo-multimedia data k nearest neighbor query
$q.loc$	Geographical location information q query objects
$q.m$	Multimedia Data Information q Query Object
$f_s(q,o)$	Spatial similarity between q and o
$f_s(q,o)$	Semantic similarity between q and o
$F_{GM}(q,o)$	q and o geo-multimedia data similarity
$H_{TDCMR}(\cdot)$	TDCMR cross-modal hash function

3.1. Problem Definition

The cross-modal hash algorithm solves the problem of uniform representation and mutual retrieval of multi-modal data through hash learning of multi-modal data. Assuming a given set of training data containing n sets $R = \{T_i, I_i, t_i\}_{i=1}^{n}$, the i-th set of training data

is composed of text modal data T_i, image modal data I_i, and category label t_i. Tuple class label subscripts constitute datasets $\Gamma = \{a_i, p_i, n_i\}_{i=1}^{m}$, and the label subscript (a_j, p_j, n_j) of the j-th group of triples indicates that the similarity between anchor sample a_i and positive sample p_i, which is higher than that between anchor sample and negative sample n_i. That is, samples with a_j and p_j have common category labels, while samples a_j and n_j do not have common category labels.

Definition 1 (Spatial similarity)**.** *Given a geo-multimedia query $q = (q.loc, q.m)$ and a space object o, then, spatial similarity between q and o is defined as ratio of euclidean space distance $\delta(q, o)$ between q and o to maximum euclidean space distance $\delta_{max}(q, O)$, which can be expressed as:*

$$f_s(q, o) = 1 - \frac{\delta(q.loc, o.loc)}{\delta_{max}(q, O)},$$

where $\delta(q.loc, o.loc)$ represents the euclidean distance between query q and spatial object o, and $\delta_{max}(q, O)$ represents the maximum euclidean distance between a query q and any spatial object in the spatial object dataset O, which can be expressed as:

$$\delta(q.loc, o.loc) = \sqrt{(q.lng - o.lng)^2 + (q.lat - o.lat)^2}$$
$$\delta_{max}(q, O) = max(\{\delta(q.loc, o.loc) | \forall o \in O\}).$$

Definition 2 (Cross-modal semantic similarity)**.** *Given a geo-multimedia Query $q = (q.loc, q.m)$ and a space object o, then, the cross-modal semantic similarity between q and o is defined as the cosine value of the TDCMR hash code and the spatial object, which can be expressed as:*

$$f_c(q, o) = \frac{\sum_{i \in q.v^T} q.v^{T(i)} \times o.v^{I(i)}}{||q.v^T|| \times ||o.v^I||}, \quad (1)$$

where v^T represents text mode TDCMR hash code, v^I represents image mode TDCMR hash code, and $||q.v^T||$ and $||o.v^I||$ are query object and space object TDCMR Hash code module.

Definition 3 (geo-multimedia Data Similarity)**.** *Given a geo-multimedia data query $q = (q.loc, q.m)$ and a geo-multimedia data object o, then, the similarity between q and o is defined as the weighted sum of spatial similarity and cross-modal semantic similarity, which can be expressed as:*

$$F_d = \lambda \cdot f_s(q, o) + (1 - \lambda) \cdot f_c(q, o), \quad (2)$$

where $f_s(q, o)$ and $f_c(q, o)$ are spatial similarity and cross-modal semantic similarity. Besides, parameter $\lambda \in [0, 1]$ is the weight factor used to balance spatial similarity and cross-modal semantic similarity. Finally, geo-multimedia data similarity score refers to the weighted score of similarity between query object and dataset object in spatial and semantic aspects, which can better meet the query processing in practical application scenarios.

Definition 4 (Spatial multimedia k nearest neighbor query)**.** *Given a Spatial Multimedia Query $q = (q.loc, q.m)$ and a space object set O, make $\forall o \in R \wedge \forall o^{'} \in (O - R)$ and $F_d(q, o) \geq F_d(q, o')$ to find a subset R having k spatial objects of O as a result of the query.*

Spatial text object query spatial image object $q_{t2i} = (q.loc, q.m^T)$ and return a result set R_{t2i} with k spatial image objects, which can be described as:

$$R_{t2i} = \{o | \forall o \in R \wedge \forall o^{'} \in (O - R_{t2i}), F_d(q_{t2i}, o) > F_d(q_{t2i}, o')\}.$$

Spatial image object query spatial text object $q_{i2t} = (q.loc, q.m^I)$ and return a result set R_{i2t} with k spatial text objects, which can be described as:

$$R_{i2t} = \{o | \forall o \in R \wedge \forall o^{'} \in (O - R_{i2t}), F_d(q_{i2t}, o) > F_d(q_{i2t}, o')\}.$$

As shown in Table 2, it gives an example of a space text object query space image object GMkNN query, as a user visited a city, because the itinerary arrangement can only selectively visit a certain scenic spot. At this point, the user can enter a text to describe the site of interest, search engine can return to the user's interest and close to the site of the image, video and other content. Users can visually browse the image or video of the scenic spot to decide whether to visit the scenic spot. In addition, there are six scenic spots, $o_1, o_2, o_3, o_4, o_5, o_6$, in the city.

Table 2. GMkNN query sample similarity score.

Query	$f_s(q,o)$	$f_c(q,o)$	$F_{GM}(q,o)$
$(q,o1)$	0.8	0.33	0.57
$(q,o2)$	0.3	0.25	0.28
$(q,o3)$	0.4	0.42	0.41
$(q,o4)$	0.4	0.92	0.66
$(q,o5)$	0.7	0.75	0.73
$(q,o6)$	0.7	0.83	0.77

For a given spatial text, query $q_{t2i} = (q.loc, q.m^T)$, and set α as 0.5. Compute the spatial similarity and cross-modal semantic similarity of query object and space object, respectively, $o_1, o_2, o_3, o_4, o_5, o_6$, and the geo-multimedia data similarity score between query object q_{t2i} and space objects $o_1, o_2, o_3, o_4, o_5, o_6$; therefore, as $k = 1$, the query q_{t2i} result of GMkNN is $\{o_2\}$, which returns the object o_2 of image content to the user.

3.2. Siamese Network and Triple Network

Siamese Network and Triple Network are members of multiple convolution neural network model. Compared with simple convolution neural network, they are composed of two or more convolution neural networks which have the same structure and shared parameters. The following mainly discusses Siamese Network and Triple Network.

3.2.1. Siamese Network

Siamese Network consists of two convolution neural networks which have the same structure and shared parameters, and it aims to map two input data to a measurable space for similarity comparison via a common function. Its objective is to minimize the similarity among the samples within the same category and maximize the similarity among the samples within the different categories. Specifically, Siamese Network searches for a set of parameters u through distance metric, such that, when T_1 and T_2 belong to the same category, the similarity is low, and, when T_1 and T_2 belong to the different categories, the similarity is high. From this, a pairwise constraint loss function is designed, which is shown in Formula (3).

$$L(T_1, T_2; m) = \frac{1}{2} \cdot I \cdot D(T_1, T_2) + \frac{1}{2} \cdot (1 - I) \cdot \{max(0, m - D(T_1, T_2))\}. \quad (3)$$

In the formula, $D(T_1, T_2)$ is defined as the square of the euclidean distance between Siamese Network outputs, as shown in Formula (4):

$$D(T_1, T_2) = ||f(T_1) - f(T_2)||_2^2. \quad (4)$$

$f(T)$ is the output of network. I represents the category relationship between sample T_1 and T_2. When samples T_1 and T_2 belong to the same category, $I = 1$, and, when samples T_1 and T_2 belong to the different categories, $I = 0$. m is a boundary value used to control the degree of the loss function.

3.2.2. Triple Network

The Triple Network model is developed based on Siamese Network, which consists of three convolution neural networks with the same structure and shared parameters. The

Triplet Network's input is a triplet that contains an anchor sample, a positive sample, and a negative sample. Generally, the anchor sample and the positive sample are sample pairs that belong to the same category or have related content, while the relationship between anchor sample and negative sample are opposite. The triplet describes the relationship between the three samples, with which the network is trained. After distance metric optimization, the distance of anchor sample is close to the positive sample and far away from the negative sample. Specially, there are N triplets defined which represent anchor sample, positive sample, and negative sample. The feature representation of the sample is obtained by the convolution neural network. According to the distance relationship between the triples constraint and the sample feature, the triples loss function is defined as below:

$$L(T_a, T_p, T_n; \alpha) = \sum_{i}^{N} max\{D(T_a, T_p) - D(T_a, T_n) + \alpha, 0\}. \quad (5)$$

In this formula, α refers to interval value of $D(T_a, T_p)$ and $D(T_a, T_n)$, similar to Formula (4), and $D(T_a, T_p)$ is defined as the square of the euclidean distance between triple network outputs.

4. Methodology

4.1. Overview of the Framework

To solve the problem of low representation ability and slow query speed in geo-multimedia data representation and query, this paper aims to narrow the cognitive gap between human and computer in multimedia data semantic understanding through a deep neural network, construct the deep cross-modal hash (Triplet-based Deep Cross-Modal Retrieval, TDCMR) network model based on triples, and encode geo-multimedia data semantically by a trained network model. Then, TDCMR Hashing Quadtree (TH-Quadtree) geo-semantic hybrid index and its query algorithm are used to search the-nearest-neighbor and semantic-related Top-k geo-multimedia data objects quickly and accurately in the massive geo-multimedia database.

4.2. Quantitative Coding of Geo-Multimedia Data Schemes

According to geo-multimedia data has the characteristics of polymorphism, heterogeneity and semantic interconnection, the traditional cross-modal hash algorithm uses low-level artificial features, and the existence of semantic gap leads to low semantic representation ability of cross-modal hash code. To better alleviate the semantic gap, this paper proposes a deep cross-modal hash algorithm based on triples TDCMR, which integrates feature learning and hash learning process through triples deep neural network, as well as designs improved triples distance constraints. The aim of this paper is to improve the semantic representation of cross-modal hash codes by forcing the same kinds of heterogeneous data close to each other in hamming space, according to the valid semantic quantization coding.

Figure 2 illustrates the proposed framework for TDCMR problem. As discussed above, the whole framework is composed of a deep feature extraction module and a hash code learning module, which is unified into an end-to-end framework. Among them, the first part is the deep feature extraction module, which consists of a multi-layer perceptron to extract text features and a deep convolutional neural network to extract image features, aiming to extract deep features of samples to alleviate the semantic gap of heterogeneous data. The second part is the hash code learning module, which constructs the semantic association between anchor sample and positive sample and negative sample through a distance learning process, and also constructs the semantic association between positive sample and anchor sample and negative sample through a distance learning process. Two distance learning processes are completed by one sample input, aiming at the heterogeneous data of the same category in hamming space being forced to approach each other.

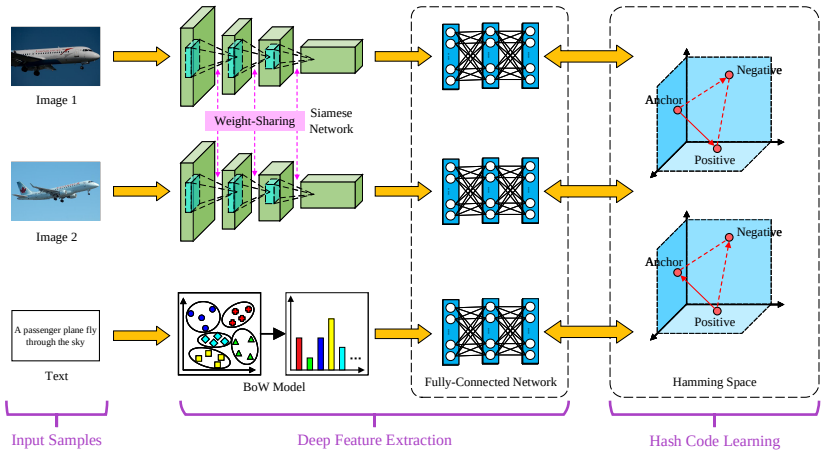


Figure 2. The framework to solve the TDCMR problem. It consists of two modules, deep feature extraction and hash code learning. To image, CNN-F transforms images from original space into feature space to hash code representation. To text, Bag of Words and MLP convert text to hash code. Then, by hash code learning, two categories of hash code are trained by triples loss to learn in hamming space.

4.3. Deep Feature Extraction

In this subsection, we employ deep convolution neural network CNN-F to extract the features of image modes and the multi-layer perceptron to extract the features of text modes. Besides, the deep feature extraction module consists of two deep neural networks.

4.3.1. CNN-F

The Deep Convolution Neural Network CNN-F network [46] consists of 5 convolutional layers and 3 fully connected layers. fc8 is a fully connected layer with a number of nodes c the length of the hash code, so as to facilitate the mapping of the image depth features extracted by the deep convolutional neural network into hash code representation. Among them, the first convolution layer conv1 uses convolution operation with step size 4, and the second convolution layer conv2 to the fifth convolution layer conv5 all use convolution operation with step size 1. In addition, maximum pooling operation in conv1, conv2, and conv5 can effectively reduce model parameters and prevent over-fitting. Similarly, the use of Dropout regularization techniques in full connection layers fc6 and fc7 can effectively prevent over-fitting.

4.3.2. MLP

Multi-layer perceptron network (MLP) consists of 3 fully connected layers. The number of nodes in layer 1 is the same as the dimension of the word bag vector input text data. The number of nodes in the layer 2 fully connected layer is set to 4096, and the number of nodes in the last layer fully connected layer is set to hash code length c, so as to facilitate the mapping of text modal depth features extracted by multi-layer perceptron network into hash code representation.

4.4. The Baseline for Triplet-Based Deep Cross-Modal Hashing

Given a set of triples$(T_{a_i}, I_{p_i}, I_{n_i})$, triplet sample distance learning goals is to close distance between anchor sample T_{a_i}, regular sample I_{p_i} and long distance between anchor sample T_{a_i} and negative sample I_{n_i}, so distance constraints of triplet samples can be defined as:

$$D(T_{a_i}, I_{p_i}) + \alpha < D(T_{a_i}, I_{n_i}), \tag{6}$$

where α is the interval value of distance $D(T_{a_i}, I_{p_i})$ between anchor sample and positive sample and the distance $D(T_{a_i}, I_{n_i})$ between anchor sample and negative sample.

The triples loss function as shown in Formula (5) can be constructed according to the triples sample distance constraint in Formula (6). During the training of the network model, it is found that, when the distance interval value α is smaller, the distance $D(T_{a_i}, I_{p_i})$ between the anchor sample T_{a_i} and the positive sample I_{p_i} is closer to the distance $D(T_{a_i}, I_{n_i})$. Although the loss function can quickly converge and close to 0, similar text and image modal samples are difficult to distinguish. When the distance interval value α is larger, the distance $D(T_{a_i}, I_{p_i})$ between the anchor sample T_{a_i} and the positive sample I_{p_i} is much smaller than the distance $D(T_{a_i}, I_{n_i})$. Similar text and image modal samples are easy to distinguish, but network models are difficult to converge.

Based on the triples sample distance constraint, the improved triples sample distance constraint is proposed. In detail, the improved triples sample distance learning goal is to construct the similarity relationship between anchor sample and positive sample and negative sample, so that the distance between anchor sample and positive sample is less than that between anchor sample and negative sample, and, at the same time, to construct the similarity relationship among positive sample, anchor sample, and negative sample, so that the distance between positive sample and anchor sample is less than that between positive sample and negative sample. By learning from two sets of sample distance relationships, heterogeneous data of the same category is forced close to each other to improve the learning ability of the network model and effectively realize that the intra-class distance is less than the inter-class distance. Improved triples sample distance constraints are formally expressed as shown in Formula (7):

$$\begin{matrix} D(T_{a_i}, I_{p_i}) + \alpha < D(T_{a_i}, I_{n_i}) \\ D(T_{p_i}, I_{p_i}) + \beta < D(T_{p_i}, I_{n_i}) \end{matrix}', \tag{7}$$

where the distance interval value α between $D(T_{a_i}, I_{p_i})$ and $D(T_{a_i}, I_{n_i})$ is a custom parameter, and the distance interval value β between $D(T_{p_i}, I_{p_i})$ and $D(T_{p_i}, I_{n_i})$ is also a custom parameter. These parameters control the distance relationship between anchor sample, positive sample, and negative sample as balance parameters.

In cross-modal hash learning, the semantic relationship between triples samples is described by the triples likelihood function. Assuming that the anchor sample is text mode and the positive sample and negative sample are image mode, the improved triples likelihood function is proposed according to the improved triples sample distance constraint in Formula (7), as shown in Formula (8):

$$p(\Gamma | B_T, B_I, B_I) = \prod_{j=1}^{m} p((a_j, p_j, n_j) | B_T, B_I, B_I), \tag{8}$$

$$p((a_j, p_j, n_j) | B_T, B_I, B_I) = \sigma(\xi_{a_j^T p_j^I} - \xi_{a_j^T n_j^I} - \alpha + \xi_{p_j^I a_j^T} - \xi_{p_j^I n_j^I} - \beta), \tag{9}$$

where $\xi_{a_j^T p_j^I} = \xi_{p_j^I a_j^T} = \frac{1}{2} G_{*a_j}^{\top} B_{I*p_j}$, $\xi_{a_j^T n_j^I} = \frac{1}{2} B_{T*a_j}^{\top} B_{I*n_j}$, $\xi_{p_j^I n_j^I} = \frac{1}{2} B_{I*p_j}^{\top} B_{I*n_j}$, $B_{T*j} = f^T T_j; u_T$, $B_{I*j} = f^I(I_j; V_I)$, B_{T*j} and B_{I*j} are the feature output of text and image modes, respectively, $B_{T*j} \in \{-1, 1\}^c$, $B_{I*j} \in \{-1, 1\}^c$, u_T and u_I are text feature extraction network and image feature extraction network, and $\sigma(T)$ denotes that the probability is calculated as a sigmoid function. α is the interval value between anchor sample and positive sample feature distance and anchor sample and negative sample feature distance, and β is the interval between the positive sample and the anchor sample and the positive sample and the negative sample.

Based on the improved triplet likelihood function, the heterogeneous association between different modal data is established by its negative logarithmic likelihood loss. The triplet loss function L_{t2i} from text mode to image mode is shown in Formula (10).

$$\begin{aligned} L_{t2i} &= -log(p(\Gamma|B_T, B_I, B_I)) \\ &= -\sum_{j=1}^{m} log(p((a_j, p_j, n_j)|B_T, B_I, B_I)) \\ &= -\sum_{j=1}^{m} (\xi_{a_j^T p_j^I} - \xi_{a_j^T n_j^I} - \alpha + \xi_{p_j^I a_j^T} - \xi_{p_j^I n_j^I} - \beta) \\ &+ \sum_{j=1}^{m} log(1 + e^{\xi_{a_j^T p_j^I} - \xi_{a_j^T n_j^I} - \alpha + \xi_{p_j^I a_j^T} - \xi_{p_j^I n_j^I} - \beta}). \end{aligned} \tag{10}$$

Similarly, the triplet loss function from image mode to text mode is shown in Formula (11).

$$\begin{aligned} L_{t2i} &= -log(p(\Gamma|B_T, B_I, B_I)) \\ &= -\sum_{j=1}^{m} log(p((a_j, p_j, n_j)|B_T, B_I, B_I)) \\ &= -\sum_{j=1}^{m} (\xi_{a_j^I p_j^T} - \xi_{a_j^I n_j^T} - \alpha + \xi_{p_j^T a_j^I} - \xi_{p_j^T n_j^I} - \beta) \\ &+ \sum_{j=1}^{m} log(1 + e^{\xi_{a_j^I p_j^T} - \xi_{a_j^I n_j^T} - \alpha + \xi_{p_j^T a_j^I} - \xi_{p_j^T n_j^I} - \beta}). \end{aligned} \tag{11}$$

Therefore, according to Formulas (10) and (11), the complete form of loss function of depth-span modal hash algorithm based on triples is shown in Formula (12):

$$\begin{aligned} min_{V,u_T,u_I} L = min_{V,u_T,u_I} L_{t2i} + L_{i2t} \\ + \gamma(||V^T - B_T||_{B_I}^2 + ||V^I - B_I||_{B_I}^2) \\ + \eta(||B_T 1||_{B_I}^2 + ||B_I 1||_{B_I}^2), \\ s.t, V^T \in \{-1, 1\}^{c \times n}, V^I \in \{-1, 1\}^{c \times n}, \end{aligned} \tag{12}$$

where B_T represents the feature vector matrix of the learned text modal data, and B_I represents the feature vector matrix of the learned image modal data; they contain the relative semantic relationship in the triple tag, and V_T, V_I, respectively, represent the text modal hash code matrix of modal and image modal data, the data feature vectors pass the semantic relationship to the corresponding hash code, where $V^T = sign(B_T)$,$V^I = sign(B_I)$. Jiang et al. [43] confirmed by a large number of experiments that better network performance can be obtained by assuming that the text mode hash code is the same as the image mode hash code during the training of the network. Therefore, the constraint condition $V = V^T = V^I$ is added on the basis of the objective loss function shown in Formula (10), and the final complete triples loss function is shown in Formula (11):

$$\begin{aligned} min_{V,u_T,u_I} L = min_{V,u_T,u_I} L_{t2i} + L_{i2t} \\ + \gamma(||V - B_T||_{B_I}^2 + ||V - B_I||_{B_I}^2) \\ + \eta(||B_T 1||_{B_I}^2 + ||B_I 1||_{B_I}^2), \\ s.t, V \in \{-1, 1\}^{c \times n}. \end{aligned} \tag{13}$$

By optimizing the loss function shown in Formula (14), the triple network can learn deep neural network parameters and hash code representation at the same time, as well as realize end-to-end learning. The first and second terms of the loss function are improved triple-negative log-likelihood loss functions. In the optimization learning process of these two terms, the cross-modal similarity of the data in the original semantic space is preserved. The third term of the loss function $\gamma(||V - B_T||_{B_I}^2 + ||V - B_I||_{B_I}^2)$ is the regularization term. By optimizing this term, the quantization error is reduced, so that the cross-modal hash

code better retains the semantic similarity in the data features. The fourth term of the loss function $\eta(||B_T1||_{B_I}^2 + ||B_I1||_{B_I}^2)$ is also a regularization term. By optimizing this term, the balance of hash code values is ensured, i.e., the number of $+1$ and -1 elements in the same position of the hash code is the same, so that each the information contained in the bit hash code is maximized.

Driven by a large number of image-text datasets, the optimized text feature extraction network parameters u_T, image feature extraction network parameters u_I, and hash code matrix V are obtained by using random gradient descent algorithm and alternating iteration strategy to quickly get and optimize the TDCMR network model; then, the network model branches are selected according to the modal types of the input data, and the deep features of the input data are extracted to obtain the cross-modal hash code.

Specifically, the optimization of the triple loss function shown in Formula (13) is a non-convex problem. Therefore, the random gradient descent algorithm simultaneously uses the alternating optimization strategy to learn parameters u_T, u_I, and the hash code matrix V, when updating one parameter, the other two parameters are fixed, the third parameter is optimized, and the optimization process is alternately carried out until the model converges or reaches the maximum number of iterations.

4.4.1. Update u_T

We learn u_T with fixed u_I and V. For each iteration, a batch-size data input network is randomly selected from the training dataset, and the back-propagation algorithm is used to learn the text features to extract the network parameters u_T. The gradient of the i-th text data object B_{T*i} to calculate the loss function is shown in Formula (14):

$$\begin{aligned}\frac{\partial L}{\partial B_{T*i}} = -\frac{1}{2}\sum_{j:(i,p_j,n_j)}^{m}(1-\sigma(2\xi_{ip_j^I}-\xi_{in_j^I}-\alpha-\xi_{p_j^Tn_j^T}-\beta))\\ (2B_{I*p_j}-B_{I*n_j})+2\gamma(B_T-V)+2\eta B_T1\end{aligned}. \tag{14}$$

Compute $\frac{\partial L}{\partial u_T}$ to update parameter u_T:

$$\frac{\partial L}{\partial u_T} = \frac{\partial L}{\partial B_{T*i}} \cdot \frac{\partial B_{T*i}}{\partial u_T}. \tag{15}$$

4.4.2. Update u_I

We learn u_I with fixed u_T and V. For each iteration, a batch-size data input network is randomly selected from the training dataset, and the back propagation algorithm is used to learn the image features to extract the network parameters u_I. The gradient of the loss function is calculated by the i-th image data object, as shown in Formula (16):

$$\begin{aligned}\frac{\partial L}{\partial B_{I*i}} = -\frac{1}{2}\sum_{j:(i,p_j,n_j)}^{m}(1-\sigma(2\xi_{ip_j^T}\\ -\xi_{in_j^T}-\alpha-\xi_{p_j^In_j^I}-\beta))\\ (2B_{T*p_j}-B_{T*n_j})\\ +2\gamma(B_I-V)+2\eta B_I1.\end{aligned} \tag{16}$$

Compute $\frac{\partial L}{\partial u_I}$ to update parameter u_I:

$$\frac{\partial L}{\partial u_I} = \frac{\partial L}{\partial B_{I*i}} \cdot \frac{\partial B_{I*j}}{\partial u_I}. \tag{17}$$

4.4.3. Update V

We have fixed parameters u_T and u_I as learning hash code matrix V. By the relation between trace and norm of matrix, to matrix N, $||N||_{B_I}^2 = tr(NN^\top) = tr(N^\top N)$. Thus, the loss function can be simplified as shown in Formula (18):

$$max_V tr(V^\top(\gamma(B_I + B_T))) \tag{18}$$

$$s.t. V \in \{-1, 1\}^{c \times n}.$$

Let $B = \gamma(B_I + B_T)$:

$$max_V \sum_{i,j} V_{ij} B_{ij} \tag{19}$$

$$s.t. V \in \{-1, 1\}^{c \times n}.$$

Keep V_{ij} and B_{ij} the same sign:

$$V = sign(\gamma(B_I + B_T)). \tag{20}$$

Given text modal data and image modal data, semantic quantization coding of text modal data and image modal data can be realized by TDCMR model. Meanwhile, semantic similarity of heterogeneous multimedia data can be measured by hamming distance. During the concrete process, the text modal data T generates semantic quantization coding v^T, and the calculation process is shown in Formula (21):

$$v^T = b^T(T) = sign(f^T(T; u_T)). \tag{21}$$

Image modal data I generate semantic quantization coding v^I, as shown in Formula (22):

$$v^I = b^I(I) = sign(f^I(I; u_I)). \tag{22}$$

As shown in the Algorithm 1, the following is our optimization procedure of the proposed TDCMR.

Algorithm 1 Optimization procedure of the proposed TDCMR

Input Textual dataset $\mathcal{T}$; Image dataset $\mathcal{I}$; Tuple label Γ;
Output Text feature extraction network parameters u_T; Image feature extraction network parameters u_I; Hash code matrix V;
Initialize parameter u_T and u_I;
The number of samples taken for each iteration $N_T = N_I = 128$;
The maximum number of iterations $t_T = \frac{n}{N_T}$ $t_I = \frac{n}{N_I}$;
for $p = 1$ to N **do**
  for $i = 1$ to t_T **do**
    Random sampling of N_T text samples from $\mathcal{T}$ to build a batch of dataset;
    Taking anchor samples from batch data to build a set of triples samples;
    For each text sample T_i, calculated $G_{*i} = f^T(T_i; u_T)$ by forward propagation;
    Gradient $\frac{\partial L}{\partial u_T}$ in Equations (3)–(10);
    Update parameters u_T by backward propagation;
  end for
  for $j = 1$ to t_I **do**
    Random sampling of N_I text samples from $\mathcal{I}$ to build a batch of dataset;
    Taking anchor samples from batch data to build a set of triples samples;
    For each text sample I_i, calculated $B_{I*i} = f^I(I_i; u_I)$ by forward propagation;
    Gradient $\frac{\partial L}{\partial u_I}$ in Equation (15);
    Update parameters u_I by backward propagation;
  end for
  Update V in Equation (18);
end for

4.4.4. Algorithm Analysis

Based on the analysis of the training effect of TDCMR algorithm, the selection of triplet samples is the key of model training effect. Give a triples sample (T_a, T_p, T_n), which is divided into the following categories:

- Simple triples: triples with a loss function value of 0. The distance of triples is satisfied $D(T_a, T_n) > D(T_a, T_p) + margin$, i.e., the distance between anchor sample T_a and positive sample T_p is less than the distance between anchor sample T_a and negative sample T_η margin, and the negative sample is easy to identify.
- Semi-difficult triples: triples with a loss function close to 0 and the distance relation of triples satisfied $D(T_a, T_p) < D(T_a, T_n) < D(T_a, T_p) + margin$, and negative sample T_n is close to anchor sample T_a, and negative sample is easy to identify.
- Difficult triples: triples with a loss function value greater than 0, the distance relation of triples satisfied $D(T_a, T_n) < D(T_a, T_p)$, and the negative sample T_n is closer to the anchor point sample T_a than the positive sample T_p, and the negative sample is difficult to identify.

The selection of triples affects the training effect of the model. Simple triples are easy to identify but cannot provide effective information for network model training. Besides, difficult triples are difficult to identify, and all difficult triples are easy to diverge the network model and seriously affect training efficiency. With the training of network model, the number of easy triples and semi-difficult triples will be much larger than the number of difficult triples, which leads to the difficulty of continuous optimization of network model in the later stage of training. Therefore, we adopt a two-stage strategy to select the three-component sample training network model. In the early stage of training, the semi-difficult triples are selected as the training data to train the network model, which makes the network model fit converge. In the later stage of training, the difficult triples are selected as the training data, and the network model is fine-tuned to obtain the optimal network model parameters and improve the training efficiency of the network model.

4.5. The TDCMR-Quadtree-Based Index Method

Figure 3 illustrates the proposed framework for index problem. The quadtree and the semantic hash table are integrated in the vertical dimension by the order of first space and then semantics. The location information is first organized according to the structure of the quadtree, then the spatial objects contained in the quadtree leaf nodes are semantically quantized by the cross-modal hash algorithm, and the hash table (Hash table) is associated to the corresponding leaf nodes according to the cross-modal hash code. The geo-semantic hybrid index TH-Quadtree is established to speed up the access to the spatial objects in the O of the geo-multimedia dataset. The structure is shown in Figure 3, and the quadtree is a tree-type index structure for accelerating spatial distance, which can organize spatial information efficiently; The TDCMR, where the cross-modal hash code describes the semantic information of geo-multimedia data, hash table supports organizing semantic information with lower storage space and search time; organizing quadtree and semantic hash table by spatial-first coupling can ensure that geo-multimedia data k nearest neighbor queries can quickly retrieve spatial objects that meet the requirements in a given spatial limitation and query semantics. The TH-Quadtree index combines the information of geo-multimedia data object space and semantics. It is a two-layer hybrid index structure, which is mainly composed of two parts: the quadtree of the spatial layer and the hash table of the semantic layer.

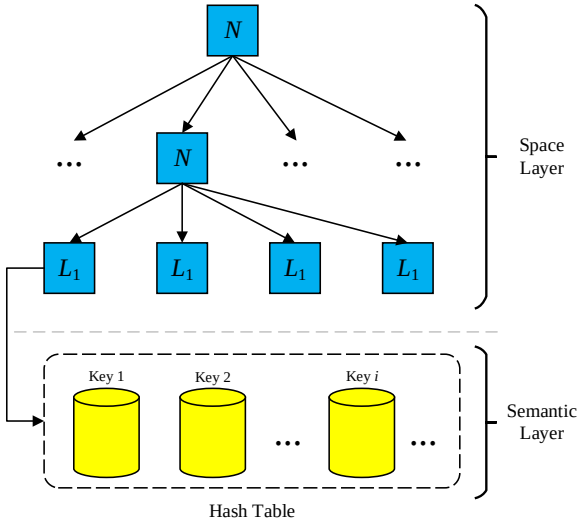


Figure 3. The framework of TDCMR-Quadtree-based index method, and it includes two layers which is space layer and semantic layer.

4.5.1. Space Layer

TH-Quadtree is a two-tier hybrid index structure that integrates the spatial and semantic layers on the vertical dimensions. In the spatial dimension, the spatial information of spatial objects is generally represented by two-dimensional latitude and longitude coordinates, which has better pruning effect than high-dimensional semantic information. Therefore, the spatial layer index is constructed by using the spatial position relation of spatial objects in geo-multimedia dataset, as the first layer of TH-Quadtree index structure. In this paper, quadtree is used to index the spatial position information of all spatial objects, which is efficient in two-dimensional spatial information organization. First, all spatial objects are regarded as the point set in the geographical space, and then each spatial object belongs to a minimum boundary rectangle MBR, i.e., each node on the quad tree, and then all the MBR are organized into different levels of tree structure according to the spatial distribution. In general, geospatial recursion is divided into hierarchical tree-type structures. Geo-multimedia data objects are all stored on each leaf node, while the root and middle nodes do not store spatial objects.

4.5.2. Semantic Layer

For each leaf node of the quadtree spatial layer, a hash table index is associated as the semantic layer of the second layer of the TH-Quadtree index structure to facilitate pruning in the semantic dimension. The semantic quantization coding of all geo-multimedia data objects in leaf nodes is obtained according to the TDCMR cross-modal hash algorithm, i.e., cross-modal hash code. Then, the uID identification code of geo-multimedia data objects is stored in a hash bucket with c bit binary encoding as key value to generate a hash table containing all spatial object semantic information of leaf nodes. The spatial objects in the hash table are located in the same hash bucket, and they have high spatial similarity, on the one hand, in the same leaf node; on the other hand, they have high semantic similarity due to the same semantic quantization coding.

4.6. TH-Quadtree-Based Nearest Neighbor Query Algorithm

The main idea of geo-multimedia data nearest neighbor query algorithm based on TH-Quadtree is: Given a query object q, the spatial object is searched orderly in the spatial layer and semantic layer. Starting with the root node of the index structure, the TH-Quadtree index structure space is traversed by TH-Quadtree index structure space according to the Spatial Best Match Proximity according to the principle of the best priority nodes of the

layer to continuously obtain the tree nodes closest to the spatial position q the query object, where the optimal spatial similarity calculation is shown in Formula (23):

$$f_{sbm}(q, N) = \lambda \cdot f_s(q, N), \tag{23}$$

where $f_s(q, N)$ stands for the spatial similarity node N and query objects q, and optimal spatial similarity $f_{sbm}(q, N)$ is the lower bound of the score of spatial object similarity o geo-multimedia data in query object q and node N. Based on the above optimal spatial similarity $f_{sbm}(q, N)$, when the query processing process accesses the leaf node, it transforms from the spatial layer search to the semantic layer search. The candidate sets related to query object semantics in the hash table associated with the leaf node are obtained quickly by Hashing Looking. Then, for the spatial object fusion spatial similarity and semantic similarity in the candidate set, the optimal spatial object update result set R is selected according to the geo-multimedia data similarity score $F_{GM}(q, o)$. During the whole search, the result set R is used maintain the traversed space object dynamically, and the current knot is formed, and a small geo-multimedia data similarity score k the fruit set is used as the upper bound of the result set, and the search is terminated when the node that has not been accessed satisfies the condition of Formula (24), and the current result set is returned as the optimal query result.

$$f_{sbm}(q, N) > D_{ub}, \tag{24}$$

where $f_{sbm}(q, N)$ is the distance lower bound of the spatial similarity between all spatial objects with N as the root node q the spatial similarity of the query object. When the distance lower bound of the spatial similarity is larger than the distance upper bound R the known result set, then, all spatial objects that are not accessed have no chance of better than the Top-k results in the current result set, then the search process terminates.

Since the spatial distance between query q and any spatial object o in node N is greater than the spatial distance from query q to node N, the spatial similarity between query q and node N will not be higher than that between query q and any spatial object o spatial similarity, i.e., $f_s(q, o) \geq f_s(q, N)$, similarly, since node N is the top element of priority queue L, the spatial similarity of query q and node N is the lower bound of the spatial similarity of all currently unvisited nodes and query q. As $f_{sbm}(q, N) > D_{ub}$, the similarity scores of all spatial objects $F_{GM}(q, o)$ that have not been accessed, and the geo-multimedia data of query q would be greater than the upper bound of distance D_{ub}, $\lambda \cdot f_s(q, o) + (1 - \lambda) \cdot f_c(q, o) > D_{ub}$. Therefore, compared with the current $Top - k$ search results, all unaccessed spatial objects have no chance to be closer to the query q, and the current result set R is the optimal solution, which can terminate the query process.

Given a geo-multimedia data k nearest neighbor query q, the distance upper bound of the result set R and the priority queue L sorted according to the spatial similarity score from small to large. In the query process, for the top element N popped by the priority queue L, the query termination condition is $\lambda \cdot f_s(q, N) > D_{ub}$.

5. Experiments

5.1. Dataset and Workload

5.1.1. Datasets

Performance of the proposed method TDCMR is evaluated on dataset MIRFlickr-25k and NUS-WIDE, and some samples of them are illustrated in Figure 4. The brief introduction of them is shown as follows.

- MIRFlickr-25k [47]. This dataset consists of 25,000 image-text pairs obtained from the Flickr website, and each pair has an image and its corresponding text labels. The dataset contains 24 manually labeled category tags, and each pair was marked with one or more category tags.
- NUS-WIDE [48]. This dataset consists of 269,648 image-text pairs obtained from Flickr website and contains 81 manually labeled category tags, and each data pair is also marked as one or more category tags.

Performance of the proposed index TH-Quadtree is evaluated on dataset real data FL and synthetic set IN. The brief introduction of them is shown as follows.

- FL. This dataset is generated by image sharing website Flickr (http://www.Flickr.com/ (accessed on 1 July 2021)), containing 1 million images with geographic location information, each containing at least one user-annotated text tag information.
- IN. This dataset is a classical image database where each node of the hierarchy is represented by at least 500 images, each concept quality controlled and manually annotated, obtaining spatial location mapping from the U.S. Place Names Commission website (http://geonames.usgs.gov(accessed on 1 July 2021)).

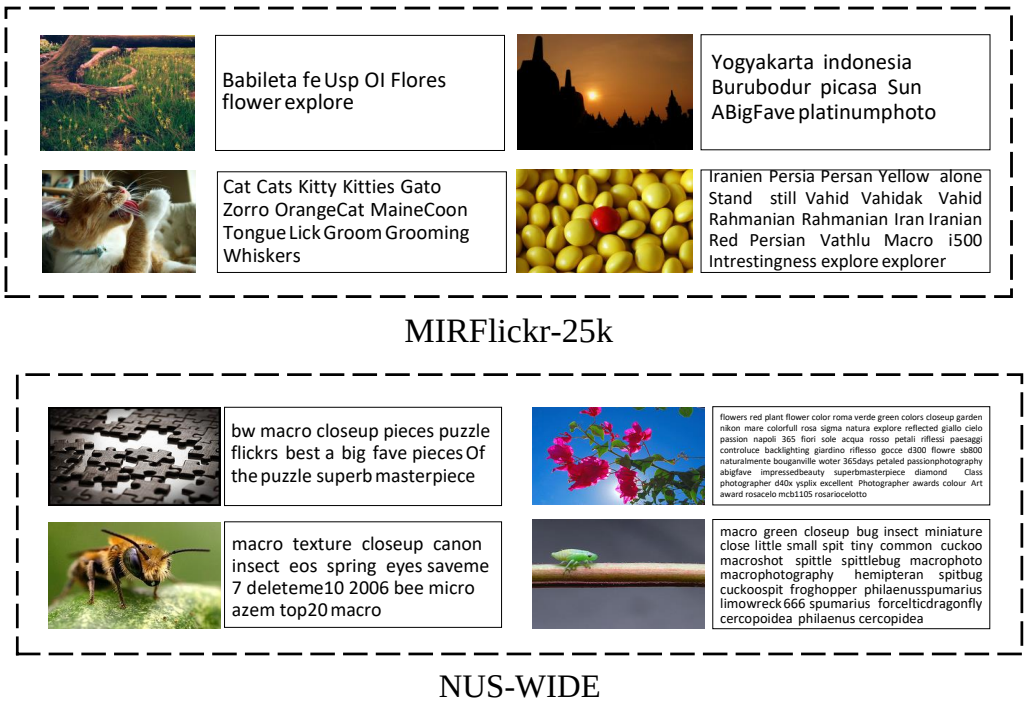


Figure 4. Some samples of MIRFlickr-25k and NUS-WIDE.

5.1.2. Workload

All the experiments are run on a PC with Intel(R) i7-6800K CPU, 64 G memory, and NVIDIA GeForce GTX 1080ti GPU, running the Ubuntu 16.04 LTS Operation System.

5.2. Settings

To evaluate the performance of the TDCMR algorithm, six cross-modal hashing algorithms are introduced for comparison in the experiments. Among these algorithms, CCA is a multivariate statistical analysis method, CVH [49], STMH [50], CMSSH [51], CMFH [52], SCM [53] and SePH [54] are shallow cross-modal hashing algorithms based on artificial features, and DCMH [43], PRDH [40], and TDH [38] are deep cross-modal hashing algorithms, which use the same network as TDCMR algorithm. We use mAP value and PR curve to quantitatively evaluate algorithm performance. We provide a Table 3 to show the benefits and drawbacks of our comparison algorithm.

In the experiment, for the MIRFlickr-25k dataset, 2000 samples are randomly extracted as the test set, and the remaining samples are used as the validation set. During the training, 10,000 samples are randomly extracted from the validation set as training data. For the NUS-WIDE dataset, 1866 samples are randomly extracted as the test set, and the remaining samples are used as the validation set. Similarly, 10,000 samples are randomly chosen from the validation set as training data. The experiment refers to the parameter settings of the paper, and sets the distance parameters α and β to half the length of the hash code. We

set the default values of balance parameters, the number of samples for each iteration 128, and the maximum number of iterations 500. If the retrieved data has the same label as the query data, it is considered to be the correct neighbor.

Table 3. The benefits and drawbacks of various methods.

Method	Benefit	Drawback
CVH	High retrieval efficiency in large-scale datasets.	Small datasets cannot give full play to the advantage of high efficiency.
STMH		
CMSSH		
CCA	Relatively effective, easy to train and implement.	It is difficult to simulate the complex correlation of cross media data, and most can only simulate two media types.
CMFH	High retrieval efficiency in large-scale datasets.	Small datasets cannot give full play to the advantage of high efficiency.
SCM	It is helpful for merging different types of information.	The optimization problem with large amount of data is a challenge.
SePH	More than two media types can be emulated.	Large time and space overhead.
PRDH	Providing more accurate information helps to improve the retrieval accuracy.	Large time and space overhead.
DCMH	Feature learning and hash code learning are combined into the same framework for the first time.	Unsupervised: cannot well overcome the barriers of different modes. Instead of manually extracting features, overcome barriers and go further.
TDH	Triplet network to process paired data and unpaired data at the same time, and learn feature expression for them.	It is difficult to adjust the training parameters due to the supervision of GAN.
TDCRM	The improved triple loss function effectively alleviates the semantic gap,Th-quadtree efficiently organizes the spatial information and semantic information of spatial multimedia data.	It cannot be directly extended to network-based studies.

5.3. Performance Evaluation

In this sectoion, the correctness of our proposed method is verified by the mAP value and PR curve, and the effectiveness of our proposed method is evaluated by the response time in the experiment.

5.3.1. Correctness Comparison

Tables 4 and 5 are the mAP values of each algorithm on the image-to-text retrieval task and the text-to-image retrieval task under different hash code lengths on the MIRFlickr-25k dataset. Tables 6 and 7 are the mAP values of each algorithm on the NUS-WIDE dataset. The best mAP values are displayed in bold font. It is observed from the mAP values in Tables 4–7 that the performance of the TDCMR algorithm outperforms the other ten cross-modal hashing algorithms.

Table 4. Performance comparison in terms of mAP scores in Image to Text on the MIRFlickr-25k dataset. The highest accuracy is shown in boldface.

Task	Method	Hash Code Length		
		16 Bit	32 Bit	64 Bit
Image To Text	CVH	0.557	0.554	0.554
	STMH	0.602	0.608	0.605
	CMSSH	0.585	0.584	0.572
	CCA	0.563	0.563	0.563
	CMFH	0.579	0.581	0.583
	SCM	0.614	0.628	0.629
	SePH	0.643	0.648	0.652
	PRDH	0.685	0.693	0.701
	DCMH	0.698	0.705	0.711
	TDH	0.702	0.711	0.718
	TDCMR	**0.711**	**0.719**	**0.727**

On the MIRFlickr-25k dataset, the performance of the TDCMR algorithm in both retrieval tasks is slightly better than the PRDH and DCMH algorithms. On the NUS-WIDE dataset, the performance of the TDCMR algorithm in both retrieval tasks is also superior to the PRDH and DCMH algorithms. The reason is that PRDH and DCMH algorithms employ pair-wise similarity constraints of category labels for hash learning without considering the relative semantic relationship between samples, which, to some extent, loses rich semantic information and results in limited retrieval accuracy. The TDCMR algorithm employs the relative semantic relationship between the three samples to build more semantic associations, while the improved triple loss function allows the hash code to retain more category information. To some extent, it overcomes the disadvantage of pairwise similarity constraint, enhances the representation ability of hash code, and, thus, has higher retrieval accuracy.

Table 5. Performance comparison in terms of mAP scores in Text to Image on the MIRFlickr-25k dataset. The highest accuracy is shown in boldface.

Task	Method	Hash Code Length		
		16 Bit	32 Bit	64 Bit
Text To Image	CVH	0.557	0.554	0.554
	STMH	0.600	0.606	0.608
	CMSSH	0.567	0.569	0.561
	CCA	0.564	0.563	0.563
	CMFH	0.578	0.579	0.581
	SCM	0.611	0.618	0.620
	SePH	0.646	0.648	0.652
	PRDH	0.728	0.732	0.739
	DCMH	0.727	0.734	0.741
	TDH	0.736	0.742	0.749
	TDCMR	**0.741**	**0.748**	**0.756**

On the MIRFlickr-25k and NUS-WIDE datasets, the retrieval performance of TDCMR, PRDH, and DCMH algorithms based on deep learning is significantly better than those based on artificial features, such as CCA, CMFH, SCM, and SePH. This is because the algorithms based on deep learning can extract deep salient features of data, and the representation ability of deep features are better than artificial features, which shows the superiority of deep neural network in saliency feature extraction.

Table 6. Performance comparison in terms of mAP scores in Image to Text on the NUS-WIDE dataset. The highest accuracy is shown in boldface.

Task	Method	Hash Code Length		
		16 Bit	32 Bit	64 Bit
Image To Text	CVH	0.400	0.392	0.386
	STMH	0.522	0.529	0.537
	CMSSH	0.511	0.506	0.493
	CCA	0.374	0.367	0.362
	CMFH	0.383	0.386	0.389
	SCM	0.489	0.494	0.499
	SePH	0.531	0.534	0.542
	PRDH	0.615	0.619	0.627
	DCMH	0.622	0.627	0.638
	TDH	0.627	0.638	0.654
	TDCMR	**0.638**	**0.651**	**0.667**

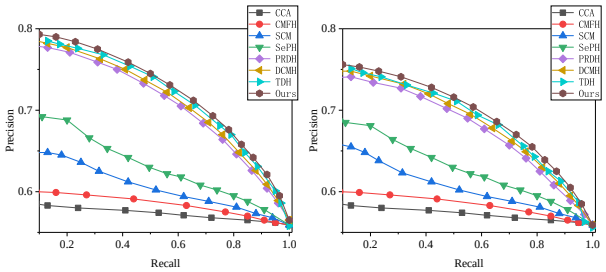
On the MIRFlickr-25k and NUS-WIDE datasets, with the increase of the hash code's length, the retrieval performance of the seven algorithms has been improved to a certain extent. The reason is that as the length of the hash code increases, the richer the semantic information contained in the code, which improves the retrieval accuracy. It should be noted that, on the one hand, the retrieval performance improves with the increase of

the code length in a certain range; on the other hand, excessive coding length leads to over-fitting and other problems, which reduces the retrieval performance.

Table 7. Performance comparison in terms of mAP scores in Text to Image on the NUS-WIDE dataset. The highest accuracy is shown in boldface.

Task	Method	Hash Code Length		
		16 Bit	32 Bit	64 Bit
Text To Image	CVH	0.372	0.366	0.363
	STMH	0.496	0.529	0.532
	CMSSH	0.449	0.389	0.380
	CCA	0.373	0.367	0.361
	CMFH	0.392	0.394	0.399
	SCM	0.460	0.466	0.469
	SePH	0.508	0.511	0.517
	PRDH	0.642	0.651	0.658
	DCMH	0.648	0.658	0.667
	TDH	0.658	0.671	0.679
	TDCMR	**0.665**	**0.678**	**0.688**

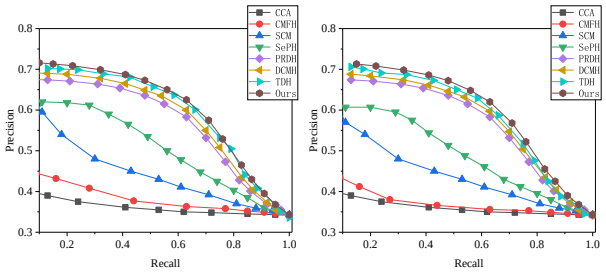
On the MIRFlickr-25k and NUS-WIDE datasets, the performance of the text-to-image retrieval task is always better than image-to-text retrieval task, probably because the hidden semantic information in image is more difficult to extract, so that the text feature extraction network can learn more information.



(**a**) (**b**)

Figure 5. Precision—recall curve on MIRFlickr-25k dataset. (**a**) T2I + 32 + MIR; (**b**) I2T + 32 + MIR.

Figure 5a,b show the accuracy recall curve of each algorithm when using 32-bit hash code on the MIRFlickr-25k dataset, while Figure 6a,b show the accuracy recall curve of each algorithm when using 32-bit hash code on NUS-WIDE dataset. According to the introduction of the accuracy recall curve, the larger the area enclosed by PR curve and coordinate axis, the better retrieval performance algorithm has. The area of TDCMR algorithm is larger than that of the other baseline methods, which is consistent with the mAP value.



(**a**) (**b**)

Figure 6. Precision—recall curve on NUS-WIDE dataset. (**a**) T2I + 32 + NUS; (**b**) I2T + 32 + NUS.

Above all, we validate the cross-modal retrieval performance of TDCMR on MIRFlickr-25k and NUS-WIDE datasets, and evaluate the retrieval accuracy through mAP value and accuracy-recall curve. The experimental results show that our proposed deep cross-modal

hash algorithm has the best performance in the cross-modal retrieval task. Therefore, TDCMR can better retain the semantic information of the original data, as well as achieve state-of-the-art retrieval performance in the field of geo-multimedia data semantic quantization coding.

5.3.2. Effectiveness Comparison

Figure 7a,b show results of this experiment in which we investigate the effect of the number of results (*k*) by varying the value *k* from 5 to 25 on dataset FL and IN. As expected, both the response time of each index increase with an increasing value of k. A larger value of *k* leads to a larger search region in query processing. Compared with traditional spatial pruning technique Quadtree, our proposed technique, TH-Quadtree, takes advantage of the semantic layer hash tables to reduce unnecessary disk access.

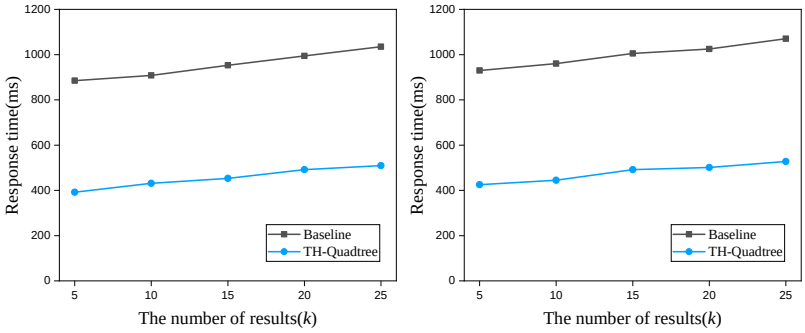


Figure 7. Varying the number of results (*k*) on FL and IN. (**a**) Varying k on FL. (**b**) Varying k on IN.

Figure 8a,b show results of this experiment in which we investigate the effect of the dataset size (n) by varying the value n from 50 K to 250 K. Similarly, when the dataset size increases, the response time of each index increases since more quadtree cells will be accessed. It is observed that the performance of TH-Quadtree is significantly superior to that of traditional quadtree.

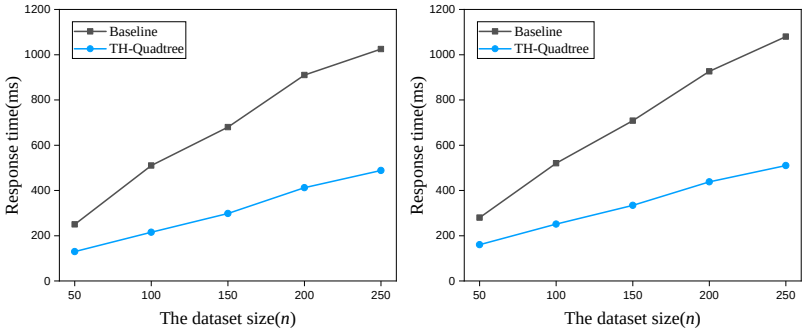


Figure 8. Varying the dataset size (n) on FL and IN. (**a**) Varying n on FL. (**b**) Varying n on IN.

6. Conclusions

To solve the problems of semantic gap and low semantic representation in the cross-modal hash method, this paper investigates a deep cross-modal hash algorithm based on TDCMR. By integrating feature extraction and hash code learning processes into an end-to-end triplet deep neural network model, sample deep features containing rich semantic information are extracted to better narrow the semantic gap. At the same time, this paper optimizes the triples sample distance relation and proposes an improved triples loss function, which makes the same heterogeneous data forced close to each other in Hamming space, and improves the semantic representation ability of the cross-modal hash code. By means of theoretical analysis and experiments, TDCMR cross-modal hash codes have better

preserved the semantic information of the original data and have better advantages in the semantic quantization coding of geo-multimedia data.

Aiming at the problems of slow speed and low efficiency in the nearest neighbor query of spatial multimedia data, a new hybrid index TH-Quadtree is proposed based on TDCMH cross-modal hash code and quadtree. TH-Quadtree efficiently organizes the spatial information and semantic information of spatial multimedia data. At the same time, based on TH-Quadtree, the nearest neighbor query algorithm is proposed to support GMkNN query, and the NE nearest neighbor expansion strategy is introduced to optimize the semantic layer search process, so as to quickly and accurately find the spatial nearest neighbor and semantically related spatial multimedia data objects in the massive spatial multimedia database. Theoretical analysis and nearest neighbor query experiments show that the nearest neighbor query algorithm of spatial multimedia data based on quadtree effectively improves the query speed.

Our future work:

(1) Semantic quantization coding based on target attention mechanism

There has always been an insurmountable semantic gap between different modal data. This paper extracts the features of image mode and text mode through a deep convolution neural network and multi-layer perceptron, which can extract the sample features with rich semantic information and make up the semantic gap to a certain extent. In the future, we try to improve and optimize the feature extraction network structure, learn the features of the target area of image or text by introducing the target attention mechanism, and obtain the data features with more significant information, and further narrow the semantic gap, so as to promote the representation ability of semantic quantitative coding.

(2) Nearest neighbor query of spatial multimedia data under road network

At present, a large number of LBS applications are based in the Euclidean space. Euclidean distance is the linear distance between any two points in space. Based on it, this paper uses Euclidean distance to measure the spatial similarity between spatial objects, and the road network distance is the road length between two points in the actual road network, which can more truly reflect people's living environment, our work tends to further expand the nearest neighbor query of spatial multimedia data to the road network environment. G-tree [55] is a road network index structure based on R-tree and preserving the spatial structure, which supports fast processing of kNN query in the road network. Therefore, how to design an efficient road network spatial semantic hybrid index structure and query algorithm based on g-tree is also content that merits being deeply studied in the future.

Author Contributions: Conceptualization, J.S. (Jiagang Song) and Y.L.; methodology, Y.L.; software, J.S. (Jiagang Song) and Y.L.; validation, J.S. (Jiagang Song), W.Y., Y.L. and J.S. (Jiayu Song); formal analysis, Y.L.; investigation, J.S. (Jiagang Song) and Y.L.; resources, Y.L.; data curation, Y.L.; writing—original draft preparation, J.S. (Jiayu Song); writing—review and editing, W.Y.; visualization, L.Z.; supervision, L.Z.; project administration, J.S. (Jiayu Song); funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by National Natural Science Foundation of China (61972203, 62072166, 61836016, 61672177) Natural Science Foundation of Jiangsu Province (BK20190442).

Institutional Review Board Statement: Not applicable.

Conflicts of Interest: The funders had no role in the design of the study.

References

1. Ouyang, J.; Liu, Y.; Shu, H. Robust hashing for image authentication using SIFT feature and quaternion Zernike moments. *Multimed. Tools Appl.* **2017**, *76*, 2609–2626. [CrossRef]
2. Zhang, C.; Lin, Y.; Zhu, L.; Zhang, Z.; Tang, Y.; Huang, F. Efficient region of visual interests search for geo-multimedia data. *Multimed. Tools Appl.* **2019**, *78*, 30839–30863. [CrossRef]
3. Xu, C.; Sun, J.; Wang, C. A novel image encryption algorithm based on bit-plane matrix rotation and hyper chaotic systems. *Multimed. Tools Appl.* **2020**, *79*, 5573–5593. [CrossRef]

4. Fang, L.; Liu, Z.; Song, W. Deep hashing neural networks for hyperspectral image feature extraction. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1412–1416. [CrossRef]
5. Cao, D.; Han, N.; Chen, H.; Wei, X.; He, X. Video-based recipe retrieval. *Inf. Sci.* **2020**, *514*, 302–318. [CrossRef]
6. Jiang, B.; Huang, X.; Yang, C.; Yuan, J. SLTFNet: A spatial and language-temporal tensor fusion network for video moment retrieval. *Inf. Process. Manag.* **2019**, *56*, 102104. [CrossRef]
7. Cao, D.; Yu, Z.; Zhang, H.; Fang, J.; Nie, L.; Tian, Q. Video-Based Cross-Modal Recipe Retrieval. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1685–1693.
8. Zhu, L.; Song, J.; Yang, Z.; Huang, W.; Zhang, C.; Yu, W. DAP^{2}CMH: Deep Adversarial Privacy-Preserving Cross-Modal Hashing. *Neural Process. Lett.* **2021**, 1–21. [CrossRef]
9. Zhang, H.L.; Huang, S. A Novel Image Authentication Robust to Geometric Transformations. In Proceedings of the 2008 Congress on Image and Signal Processing, Sanya, China, 27–30 May 2008; Volume 2, pp. 654–658.
10. Cao, D.; Chu, J.; Zhu, N.; Nie, L. Cross-modal recipe retrieval via parallel-and cross-attention networks learning. *Knowl.-Based Syst.* **2020**, *193*, 105428. [CrossRef]
11. Wu, L.; Wang, Y.; Yin, H.; Wang, M.; Shao, L. Few-shot deep adversarial learning for video-based person re-identification. *IEEE Trans. Image Process.* **2019**, *29*, 1233–1245. [CrossRef]
12. Liu, Y.; Qin, Z.; Liao, X.; Wu, J. Cryptanalysis and enhancement of an image encryption scheme based on a 1-D coupled Sine map. *Nonlinear Dyn.* **2020**, *100*, 2917–2931. [CrossRef]
13. Deng, G.; Xu, C.; Tu, X.; Li, T.; Gao, N. Rapid image retrieval with binary hash codes based on deep learning. In Proceedings of the Third International Workshop on Pattern Recognition, Jinan, China, 26–28 May 2018; International Society for Optics and Photonics: Bellingham, WA, USA; Volume 10828, p. 1082813.
14. Zhu, L.; Zhang, C.; Song, J.; Liu, L.; Zhang, S.; Li, Y. Multi-Graph Based Hierarchical Semantic Fusion for Cross-Modal Representation. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
15. Zheng, W.; Zhu, X.; Zhu, Y.; Zhang, S. Robust Feature Selection on Incomplete Data. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 3191–3197.
16. Zhang, C.; Song, J.; Zhu, X.; Zhu, L.; Zhang, S. HCMSL: Hybrid Cross-modal Similarity Learning for Cross-modal Retrieval. *ACM Trans. Multimed. Comput. Commun. Appl.* **2021**, *17*, 1–22. [CrossRef]
17. Jiang, B.; Huang, X.; Yang, C.; Yuan, J. Cross-modal video moment retrieval with spatial and language-temporal attention. In Proceedings of the 2019 on International Conference on Multimedia Retrieval, Ottawa, ON, Canada, 10–13 June 2019; pp. 217–225.
18. Hu, R.; Zhu, X.; Zhu, Y.; Gan, J. Robust SVM with adaptive graph learning. *World Wide Web* **2020**, *23*, 1945–1968. [CrossRef]
19. Zhu, L.; Song, J.; Zhu, X.; Zhang, C.; Zhang, S.; Yuan, X. Adversarial Learning-Based Semantic Correlation Representation for Cross-Modal Retrieval. *IEEE Multimed.* **2020**, *27*, 79–90. [CrossRef]
20. Zhu, L.; Song, J.; Wei, X.; Yu, H.; Long, J. CAESAR: Concept augmentation based semantic representation for cross-modal retrieval. *Multimed. Tools Appl.* **2020**, 1–31. [CrossRef]
21. Zhu, X.; Zhu, Y.; Zheng, W. Spectral rotation for deep one-step clustering. *Pattern Recognit.* **2020**, *105*, 107175. [CrossRef]
22. Wu, L.; Wang, Y.; Gao, J.; Wang, M.; Zha, Z.J.; Tao, D. Deep Coattention-Based Comparator for Relative Representation Learning in Person Re-Identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 722–735. [CrossRef] [PubMed]
23. Adam, M.; Tomáek, P.; Lehejek, J.; Trojan, J.; Jnek, T. The Role of Citizen Science and Deep Learning in Camera Trapping. *Sustainability* **2021**, *13*, 10287. [CrossRef]
24. Franzen, M.; Kloetzer, L.; Ponti, M.; Trojan, J.; Vicens, J. Machine Learning in Citizen Science: Promises and Implications. In *The Science of Citizen Science*; Springer: Cham, Switzerland, 2021.
25. Liu, Y.; Xin, G.; Xiao, Y. Robust Image Hashing Using Radon Transform and Invariant Features. *Radioengineering* **2016**, *25*, 556–564. [CrossRef]
26. Ouyang, J.; Wen, X.; Liu, J.; Chen, J. Robust hashing based on quaternion zernike moments for image authentication. *ACM Trans. Multimed. Comput. Commun. Appl.* **2016**, *12*, 1–13. [CrossRef]
27. Zhang, H.; Xiong, C.; Geng, G. Content based image hashing robust to geometric transformations. In Proceedings of the 2009 Second International Symposium on Electronic Commerce and Security, Nanchang, China, 22–24 May 2009; Volume 2; pp. 105–108.
28. Wang, Y. Survey on deep multi-modal data analytics: Collaboration, rivalry and fusion. *arXiv* **2020**, arXiv:2006.08159.
29. Zhang, C.; Zhang, Y.; Zhang, W.; Lin, X. Inverted linear quadtree: Efficient top *k* spatial keyword search. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1706–1721. [CrossRef]
30. Cong, G.; Jensen, C.S.; Wu, D. Efficient retrieval of the top-k most relevant spatial web objects. *Proc. VLDB Endow.* **2009**, *2*, 337–348. [CrossRef]
31. Zhang, D.; Chan, C.Y.; Tan, K.L. Processing spatial keyword query as a top-k aggregation query. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Australia, 6–11 July 2014; pp. 355–364.
32. Rocha-Junior, J.B.; Vlachou, A.; Doulkeridis, C.; Nørvåg, K. Efficient processing of top-k spatial preference queries. *Proc. VLDB Endow.* **2010**, *4*, 93–104. [CrossRef]

33. Zhang, D.; Tan, K.L.; Tung, A.K. Scalable top-k spatial keyword search. In Proceedings of the 16th International Conference on Extending Database Technology, Genoa, Italy, 18–22 March 2013; pp. 359–370.
34. Zhu, L.; Song, J.; Yu, W.; Zhang, C.; Zhang, Z. Reverse Spatial Visual Top-k Query. *IEEE Access* **2020**, *8*, 21770–21787. [CrossRef]
35. Zhang, C.; Cheng, K.; Zhu, L.; Chen, R.; Zhang, Z.; Huang, F. Efficient continuous top-k geo-image search on road network. *Multimed. Tools Appl.* **2019**, *78*, 30809–30838. [CrossRef]
36. Zhang, C.; Zhu, L.; Zhang, S.; Yu, W. TDHPPIR: An Efficient Deep Hashing Based Privacy-Preserving Image Retrieval Method. *Neurocomputing* **2020**, *406*, 386–398. [CrossRef]
37. Liong, V.E.; Lu, J.; Tan, Y.; Zhou, J. Cross-Modal Deep Variational Hashing. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 4097–4105. [CrossRef]
38. Deng, C.; Chen, Z.; Liu, X.; Gao, X.; Tao, D. Triplet-Based Deep Hashing Network for Cross-Modal Retrieval. *IEEE Trans. Image Process.* **2018**, *27*, 3893–3903. [CrossRef] [PubMed]
39. Wu, L.; Wang, Y.; Gao, J.; Li, X. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Trans. Multimed.* **2018**, *21*, 1412–1424. [CrossRef]
40. Yang, E.; Deng, C.; Liu, W.; Liu, X.; Tao, D.; Gao, X. Pairwise Relationship Guided Deep Hashing for Cross-Modal Retrieval. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 1618–1625.
41. Zhu, L.; Long, J.; Zhang, C.; Yu, W.; Yuan, X.; Sun, L. An Efficient Approach for Geo-Multimedia Cross-Modal Retrieval. *IEEE Access* **2019**, *7*, 180571–180589. [CrossRef]
42. Wang, Y.; Wu, L.; Lin, X.; Gao, J. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 4833–4843. [CrossRef] [PubMed]
43. Jiang, Q.Y.; Li, W.J. Deep cross-modal hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3232–3240.
44. Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; Tao, D. Self-supervised adversarial hashing networks for cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–22 June 2018; pp. 4242–4251.
45. Zhu, L.; Yu, W.; Zhang, C.; Zhang, Z.; Huang, F.; Yu, H. SVS-JOIN: Efficient Spatial Visual Similarity Join for Geo-Multimedia. *IEEE Access* **2019**, *7*, 158389–158408. [CrossRef]
46. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
47. Huiskes, M.J.; Lew, M.S. The MIR flickr retrieval evaluation. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, 30–31 October 2008; pp. 39–43.
48. Chua, T.S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. 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, Santorini Island, Greece, 8–10 July 2009; pp. 1–9.
49. Kumar, S.; Udupa, R. Learning hash functions for cross-view similarity search. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.
50. Wang, D.; Gao, X.; Wang, X.; He, L. Semantic topic multimodal hashing for cross-media retrieval. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
51. Bronstein, M.M.; Bronstein, A.M.; Michel, F.; Paragios, N. 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, San Francisco, CA, USA, 13–18 June 2010; pp. 3594–3601.
52. Ding, G.; Guo, Y.; Zhou, J. Collective matrix factorization hashing for multimodal data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 2075–2082.
53. Zhang, D.; Li, W.J. Large-scale supervised multimodal hashing with semantic correlation maximization. In Proceedings of the AAAI, Québec City, QC, Canada, 27–31 July 2014; Volume 1, p. 7.
54. Lin, Z.; Ding, G.; Hu, M.; Wang, J. Semantics-preserving hashing for cross-view retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3864–3872.
55. Zhong, R.; Li, G.; Tan, K.L.; Zhou, L. G-tree: An efficient index for KNN search on road networks. In Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013.