This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Submitted May 21, 2023

*Digital Object Identifier*

# Transformer-based Discriminative and Strong Representation Deep Hashing for Cross-Modal Retrieval

**SUQING ZHOU[1], YU HAN[2], NING CHEN[3], SIYU HUANG[4], KOSTROMITIN KONSTANTIN IGOREVICH[5], JIA LUO[6], AND PEIYING ZHANG[3,7,8]**

[1]Internet of Things and Artificial Intelligence College, Fujian Polytechnic of Information Technology, Fuzhou 350003, China
[2]China Mobile Group Shandong Co., Ltd., Jinan 265400, China
[3]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China
[4]Xiongan Institute of Innovation, Chinese Academy of Sciences, Baoding 071702, China
[5]Department of Physics of Nanoscale Systems, South Ural State University, Chelyabinsk 454080, The Russian Federation
[6]College of Economics and Management, Beijing University of Technology, Beijing 100124, China
[7]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[8]Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science, Jinan 250353, China

Corresponding author: Yu Han (e-mail: hanyu@sd.chinamobile.com).

This work is partially supported by the Natural Science Foundation of Shandong Province under Grant ZR2023LZH017 and ZR2022LZH015, partially supported by the Industry-university Research Innovation Foundation of Ministry of Education of China under Grant 2021FNA01001, partially supported by the Key Funding from National Natural Science Foundation of China under Grant 92067206.

**ABSTRACT** Cross-modal hashing retrieval has attracted extensive attention due to its low storage requirements as well as high retrieval efficiency. In particular, how to more fully exploit the correlation of different modality data and generate a more distinguished representation is the key to improving the performance of this method. Moreover, Transformer-based models have been widely used in various fields, including natural language processing, due to their powerful contextual information processing capabilities. Based on these motivations, we propose a Transformer-based Distinguishing Strong Representation Deep Hashing (TDSRDH). For text modality, since the sequential relations between words imply semantic relations that are not independent relations, we thoughtfully encode them using a transformer-based encoder to obtain a strong representation. In addition, we propose a triple-supervised loss based on the commonly used pairwise loss and quantization loss. The latter two ensure the learned features and hash-codes can preserve the similarity of the original data during the learning process. The former ensures that the distance between similar instances is closer and the distance between dissimilar instances is farther. So that TDSRDH can generate more discriminative representations while preserving the similarity between modalities. Finally, experiments on the three datasets *MIRFLICKR-25K*, *IAPR TC-12*, and *NUS-WIDE* demonstrated the superiority of TDSRDH over the other baselines. Moreover, the effectiveness of the proposed idea was demonstrated by ablation experiments.

**INDEX TERMS** Cross-Modal Retrieval; Deep Hashing; Transformer; Strong Representation; Discriminative Representation

## I. INTRODUCTION

WITH the development of communications technology and the widespread popularity of smart devices, massive multi-modal data such as text, image, video, and audio have exploded on the Internet [1], [2]. It is noteworthy that these multi-modal data are heterogeneous and multiple in form, but interrelated in semantics [3]–[6]. Therefore, cross-modal retrieval has attracted extensive attention. However, the different structures and feature distributions of different modal data make it impossible to compare multimodal data directly. For this reason, how to effectively mine the semantic consistency and correlation between different modalities is still a challenging problem [7].

Moreover, considering the high dimensionality and massiveness of multi-modal data, hash learning is used with real-time cross-modal retrieval due to its low storage requirements



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



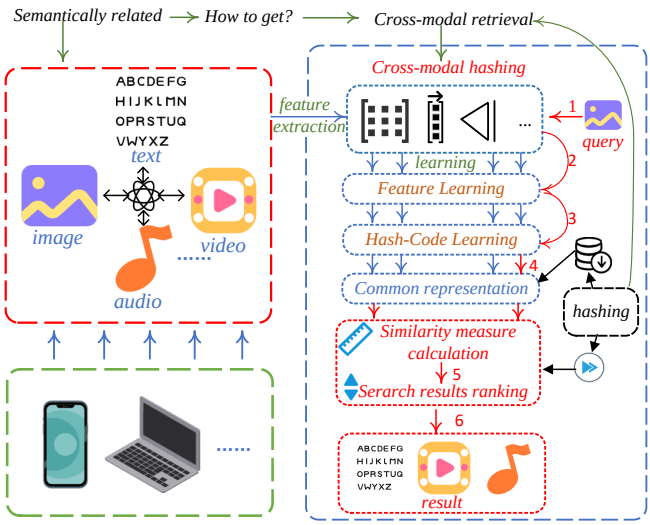



**FIGURE 1.** Cross-modal hashing retrieval. The image is used as a query to describe the process of cross-modal retrieval. After two main processes of feature learning and hash-code learning, the modality data is mapped into the hash-code in the public Hamming space. The hash-code of the modality to be queried is measured for similarity with the hash-code in the database, and the query result is obtained after sorting.

and fast retrieval speed [8]–[11]. Hashing aims to map the modal data of the high-dimensional feature space into a compact binary hash-code in the Hamming space [12], where the semantically similar instances in the original space should also have similarities in the Hamming space, that is the principle of *preserving similarity* [13]. Therefore, in cross-modal hashing, how to effectively use semantic consistency and relevance as well as generate more discriminative hash-codes is still a challenging problem, while preserving the original semantic similarity [14]. In this work, the loss function we proposed can effectively utilize semantic information while generating more discriminative representations.

According to whether to use semantic labels, the traditional cross-modal hashing methods are divided into supervised learning [15], [16] and unsupervised learning [13], [17]. Compared with the latter, supervised learning uses semantic labels to establish the correlation between data of different modalities, which brings better performance [18]–[20]. Whereas, such methods rely on hand-made features extracted through shallow structures, which separate the feature extraction process from the hash-code learning process and reduce the effectiveness of the learned code. Fig. 1 illustrates the process of the cross-modal hashing method. Based on the success of deep learning, compared with the shallow structure, the feature representation extracted by deep learning technology is more discriminative and effective [21]. Accordingly, it is widely used in cross-modal hashing. Most methods are composed of feature learning and hash-code learning, both of which play a vital role in improving the performance of the model. Unfortunately, numerous methods pay more attention to the hash-code learning part and pay less attention to the feature learning part. Only the weak representation of the feature is obtained, while the extraction of the feature and context information is slightly rough. Especially for the text modalities, most methods merely use multi-layer fully connected layers for feature extraction [22]–[24]. Subsequently, inspired by LSTM [25], many works [26] use LSTM to extract text features, and the result is significantly improved. For the text, the order between words is exceedingly important [27]. The majority of the previous methods, using bag-of-words (BOW) to process text data only consider the desired words, but neglect the order relationship between words. In this way, they get discrete and sparse features in most cases. It assumes that words are independent, on the contrary, words affect each other [28]. Enlightened by Transformer [29] and its success in the field of natural language processing [30], we utilize a transformer-based encoder for text modal. In the process, word embedding is employed to process original text data and position information is embedded, which fully preserves the semantic information implicit in the order relationship of the text data. Furthermore, we also consider contextual information to effectively use semantic information to obtain a strong representation of features.

In this paper, we propose a novel deep cross-modal hashing framework, called Transformer-based Discriminative and Strong Representation Deep Hashing for Cross-Modal Retrieval (TDSRDH). The main *contributions* of TDSRDH can be summarized as follows:

- TDSRDH seamlessly integrates feature learning and hash-code learning, each end of which is based on the high-performance encoder to obtain a strong representation of the features, where we innovatively utilize a transformer-based encoder to encode the text modality.
- We propose a triple-supervised loss and combine it with pairwise loss and quantization loss, so that TDSRDH can generate more discriminative representations while preserving the similarity.
- Extensive experiments have been conducted on three widely used datasets (*MIRFLICKR-25K*, *IAPR TC-12*, and *NUS-WIDE*), and the results demonstrate that TDSRDH can outperform other baselines by $1\%$ to $3\%$ to achieve promising performance.
- Eventually, ablation experiments have been performed, verifying our viewpoints can bring about $2\%$ and $1\%$ improvement respectively, proving their effectiveness.

The rest of the paper is organized as follows. The related work is reviewed in Section II. Section III manifests the problem definition of this paper. The proposed TDSRDH framework is presented in Section IV, including detailed information on the model and the whole process of learning. In Section V, the results of our experiment are provided. Finally, in Section VI, we summarize our work.

## II. RELATED WORK

According to the way of feature extraction, cross-modal hashing can be divided into shallow structure-based methods and deep structure-based methods. Similarly, according to



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



whether to use semantic label information, it can be divided into two categories: unsupervised methods and supervised methods.

### A. SHALLOW STRUCTURE-BASED

#### 1) unsupervised

The unsupervised methods preserve the similarity between heterogeneous data through the co-occurrence information of different modal data without the supervised information such as semantic labels, then project the data into the joint embedding space [31]. To explore the correlation of multi-modal data from different media, Inter-Media Hashing (IMH) [32] obtains a more effective hash-code by calculating the similarity between the modal and within the modal in the public Hamming space. On this basis, Collective Matrix Factorization Hashing (CMFH) [13] firstly proposes the use of matrix factorization technology to learn cross-view hash functions and has improved retrieval accuracy by merging multiple view information sources. However, owing to the semantic label information have not being used, there is still room for further improvement of the retrieval results.

#### 2) supervised

The supervised methods utilize semantic labels as supervised information and learn more discriminative hash representations [33]. Semantic Correlation Maximization (SCM) [15] seamlessly integrates semantic information into the hash-code learning process of large-scale data modeling and constructs the similarity matrix through semantic labels. Semantics-Preserving Hashing (SePH) [16] utilizes the semantic similarity of the data as supervised information and minimizes the Kulback-Leibler divergence in the Hamming space to learn hash-codes. Furthermore, considering that data from different modalities may not have a one-to-one correspondence, equal-length hash-codes will sacrifice the scalability of their representation. Matrix Tri-Factorization Hashing (MTFH) [34] proposes to learn hash-codes of different lengths for the first time, which can work seamlessly in various retrieval tasks, including paired or unpaired, and hash coding environments of equal or unequal length. In order to make full use of the inherent correlation between different modalities, Asymmetric Supervised Consistent and Specific Hashing (ASCSH) [35] decomposes the mapping matrix into consistent and modality-specific parts, which effectively improves the discrimination of hash-codes. The addition of semantic supervised information makes the learned hash-code more discriminative, which improves the performance of the model. Nevertheless, these methods are based on a shallow structure, which limits the full exploration of semantic information between modalities and reduces the effectiveness of the learned hash-codes.

### B. DEEP STRUCTURE-BASED

On account of the features extracted by deep learning being more discriminative and effective, as well as powerful non-linear representation capabilities [36], it plays an increasingly important role in cross-modal hashing. The most representative one is Deep Cross-Modal Hashing (DCMH) [22], which first proposes the combination of deep learning and hash-code learning to solve the cross-modal retrieval problem and proves its feasibility. Many pieces of research based on deep hashing have been launched one after another. On this basis, considering that DCMH lacks an in-depth exploration of semantic information, Pairwise Relationship guided Deep Hashing (PRDH) [24] introduces decorrelation constraints and different types of pairwise constraints to preserve the similarity of hash-codes while enhancing the distinguishing ability of each hash bit. Semantic Deep Cross-Modal Hashing (SDCH) [37] adds semantic branches and employs semantic labels to improve the hash-code learning part, so that the learned hash-code retains the semantic information of the original data. Deep Supervised Cross-Modal Retrieval (DSCMR) [38] minimizes the discrimination loss of the sample to learn the discriminative features of the supervised model. At the same time, it uses the weight-sharing strategy to eliminate cross-modal differences, so that the paired label information and classification information are fully utilized, and the heterogeneous differences between modal data are better bridged. In order to preserve the multi-level semantic similarity between multi-label objects, Ranking-based Deep Cross-Modal Hashing (RDCMH) [39] seamlessly integrates deep feature learning and hashing based on semantic sorting, and introduces supervised semantic metrics induced by label and feature information to utilize labeled and unlabeled data. It has achieved satisfactory results in the semi-supervised environmental environment. Inspired by the adversarial network [40], Self-Supervised Adversarial Hashing (SSAH) [41] utilizes two adversarial networks to learn high-dimensional features and hash codes and combines self-supervised semantic learning with adversarial learning to maximize semantic relevance and consistency of feature distribution. After that, combined with the work of attention [29], Attention-Aware Deep Adversarial Hashing (AADAH) [23] puts forward an adversarial network combining attention mechanism and hashing, which selectively focuses on the information part of multi-modal data to enhance the measurement of content similarity. This method has brought substantial progress. Multi-pathway Generative Adversarial Hashing (MGAH) [42] makes full use of the unsupervised representation learning capabilities of GAN networks to further explore the potentially complex structure of cross-modal data. To fully explore the semantic consistency and correlation between different modalities, Multi-Task Consistency-Preserving Adversarial Hashing (CPAH) [43] seamlessly integrates multi-modal semantic consistency learning and hashing learning. It also proposes a consistency refined module and a multi-task adversarial learning module for modal representation and retention of semantic consistency information. To solve the problem of inaccurate similarity in unsupervised cross-modal hashing, Deep Graph-neighbor Coherence Preserving



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



Network (DGCPN) [44] combines the graph model as well as integrates the data and the relevant information between its neighbors to explore the consistency of graph-neighbors to solve this dilemma. To better integrate contrastive learning with unsupervised hashing, Unsupervised Contrastive Cross-Modal Hashing (UCCH) [45] proposes a novel momentum optimizer and introduces a cross-modal ranking learning loss (CRL). The global comparison strategy it uses obtains better performance and is one of the first contrastive hashing methods. To address the problem that labels similarity and relative ranking between neighboring instances have been neglected in most of the existing works, Deep Cross-modal Hashing based on Semantic Consistency Ranking (DCH-SCR) [46] integrates the semantic similarity in features and labels for supervision. It proposes a rank alignment loss function for narrowing the semantic gap between modalities and optimizes compact hash codes. In order to more fully exploit the semantic correlations among modalities, multi-label semantic supervised graph attention hashing (MS$^2$GAH) [11] proposes a deep hashing method combined with graph attention networks. Specifically, it uses graph networks to construct graph features and combines attention mechanisms to assign importance weights to different neighbors. In addition, multi-label annotations are combined to construct semantic correlations in a fine-grained manner, which leads to a more accurate modal relationship. In addition, to preserve more semantic information, Yang *et al.* use discriminative labels and propose High-level Semantic Similarity Analysis Hashing (HSSAH) [47]. Unlike other works that use the semantic similarity matrix, this work constructs an asymmetric high-level semantic similarity matrix. To combine the semantic information in labels more efficiently and reduce the high computational requirements of similarity calculation, Fast Asymmetric Discrete Cross-Modal Hashing (FADCH) [48] utilizes matrix decomposition to construct the public semantic space. In addition, it uses label construction semantics to semantically align with the public semantic space to make full use of the semantic information. And, it designs an asymmetric strategy to associate semantics with hash-code. To address how existing semi-supervised cross-modal hashing methods can adequately explore modal complementarity and modal correlation information, the Modality-specific and Cross-Modal Graph Convolutional Networks (MCGCN) [49] proposes a three-channel based approach (two modal channels and one cross-modal channel) using graph convolutional networks. It also designs a discriminative model and guides the training process through an adversarial scheme to narrow the modal gap.

On the contrary, the majority of these methods ignore the feature extraction part to a certain extent. In particular, text modalities use scattered and sparse representations. Most methods merely employ multi-layer fully connected layers for feature extraction, without considering their semantic relationships, and just gain weak representations of features. For text modalities, TDSRDH utilizes a transformer-based encoder for text modal feature extraction to effectively make use of the semantic information and obtain a strong representation of features. At the same time, how to generate more discriminative hash-codes is an ongoing challenge to be solved. We propose a triple-supervised loss and combine pairwise loss and quantization loss to generate more discriminative hash-codes while preserving similarity.

## III. PROBLEM DEFINITION

### A. NOTATION

**TABLE 1.** Definition of related notations

| Description | Notation |
|---|---|
| scalar (lowercase) | $a$ |
| vector (boldface lowercase) | $\boldsymbol{a}$ |
| matrix (boldface uppercase) | $\boldsymbol{A}$ |
| the $i$-th row of $\boldsymbol{A}$ | $\boldsymbol{A}_{i*}$ |
| the $j$-th column of $\boldsymbol{A}$ | $\boldsymbol{A}_{*j}$ |
| the element in $i$-th row and $j$-th column of $\boldsymbol{A}$ | $\boldsymbol{A}_{ij}$ |
| the transpose of $\boldsymbol{A}$ | $\boldsymbol{A}^T$ |
| the Frobenius norm of $\boldsymbol{A}$ | $\|\boldsymbol{A}\|_F$ |
| the trace of $\boldsymbol{A}$ | $tr(\boldsymbol{A})$ |
| element sign function | $\operatorname{sign}(a) = \begin{cases} 1, & a \geq 0, \\ -1, & a < 0. \end{cases}$ |

The related definitions of notations used in this work are shown in Table 1.

### B. PROBLEM FORMULATION

TDSRDH can be easily extended to multiple modalities, such as video, audio, and so on. Without loss of generality, we only use image and text to reveal our method. The related data notations of TDSRDH are shown in Table 2. Here, $\boldsymbol{L}$ is composed of elements of 0 or 1 where 1 means that the label is contained, and 0 means that the label is not contained. Accordingly, as long as the $\boldsymbol{V}_{i*}$ and the $\boldsymbol{W}_{j*}$ share at least

**TABLE 2.** Data notations of TDSRDH.

| Notation | Description |
|---|---|
| $n$ | the number of training instances(image-text pairs) |
| $m$ | the number of category of labels |
| $k$ | the length of hash-code |
| $d_v$ | the dimension of each image instance |
| $d_w$ | the dimension of each text instance |
| $\boldsymbol{V} \in \mathbb{R}^{n \times d_v}$ | the data of image modality |
| $\boldsymbol{W} \in \mathbb{R}^{n \times d_w}$ | the data of text modality |
| $\boldsymbol{V}_{i*}$ | the original data of the image $i$ |
| $\boldsymbol{W}_{i*}$ | the original data of the text $i$ (related to $\boldsymbol{V}_{i*}$) |
| $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ | the cross-modal similarity matrix |
| $\boldsymbol{L} \in \mathbb{R}^{n \times m}$ | the multi-label matrix |
| $\boldsymbol{L}_{i*} \in \{0, 1\}$ | the common multi-labels of $\boldsymbol{V}_{i*}$ and $\boldsymbol{W}_{i*}$ |



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



one label, we consider them similar. In this case, $\boldsymbol{S}_{ij} = 1$, otherwise, they have different class labels, $\boldsymbol{S}_{ij} = 0$.

Given the training data information $\boldsymbol{V}$, $\boldsymbol{W}$, and $\boldsymbol{L}$, our goal is to learn the hash function of the image and text modal respectively, and get the hash-code: $\boldsymbol{B}^{(v)} = f(\boldsymbol{V}) \in \{-1, +1\}^{n \times k}$ and $\boldsymbol{B}^{(w)} = g(\boldsymbol{W}) \in \{-1, +1\}^{n \times k}$, where $\boldsymbol{B}^{(v)}$ and $\boldsymbol{B}^{(w)}$ are the hash-codes obtained after $\boldsymbol{V}$ and $\boldsymbol{W}$ are converted by $f(\bullet)$ and $g(\bullet)$ respectively. The point to note is that $\boldsymbol{S}$ is calculated by $\boldsymbol{L}$ and used for the supervision of two hash functions to preserve the cross-modal similarity in $\boldsymbol{S}$. The similarity of hash-codes is measured by the Hamming distance which can be formulated as:

$$d_{ij} = \frac{1}{2}\left(k - \boldsymbol{B}_{i*}^{(v)}\left(\boldsymbol{B}_{j*}^{(w)}\right)^{T}\right). \tag{1}$$

It is worth noting that here we measure the similarity between hash-codes in the form of the inner product. Given $\boldsymbol{S}$, the probability of $\boldsymbol{S}$ under the conditions $\boldsymbol{B}_{i*}^{(v)}$ and $\boldsymbol{B}_{j*}^{(w)}$ can be expressed as:

$$p\left(\boldsymbol{S}_{ij} \mid \boldsymbol{B}_{i*}^{(v)}, \boldsymbol{B}_{j*}^{(w)}\right) = \begin{cases} \sigma\left(\Theta_{ij}\right), & S_{ij} = 1, \\ 1 - \sigma\left(\Theta_{ij}\right), & S_{ij} = 0, \end{cases} \tag{2}$$

where $\Theta_{ij} = \frac{1}{2}\boldsymbol{B}_{i*}^{(v)}\left(\boldsymbol{B}_{j*}^{(w)}\right)^{T}$ and $\sigma\left(\Theta_{ij}\right) = \frac{1}{1+e^{-\Theta_{ij}}}$. More specifically, if $\boldsymbol{S}_{ij} = 1$, the inner product between the hash-codes $\boldsymbol{B}_{i*}^{(v)}$ and $\boldsymbol{B}_{j*}^{(w)}$ should be large, the Hamming distance $d_{ij}$ should be small, and the probability $p$ should be high. Otherwise, if $\boldsymbol{S}_{ij} = 0$, the inner product should be small, the Hamming distance should be large and the probability should be low.

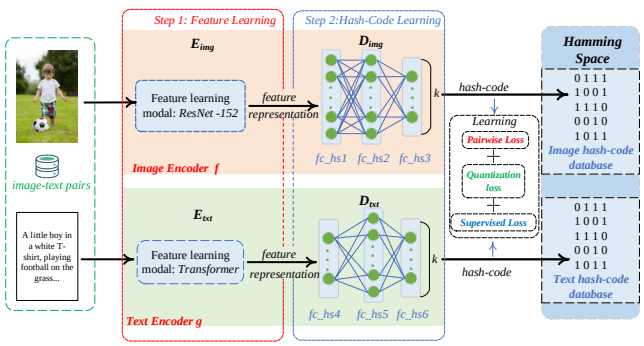



**FIGURE 2.** The framework of TDSRDH. TDSRDH is an end-to-end framework with two encoder networks: Image Encoder $f$ and Text Encoder $g$, and each contains two models: feature learning model $E$ and hash-code learning model $D$. $E_{img}$ and $E_{txt}$ are image and text modal feature learning models, respectively. "$fc_hs1 \rightarrow fc_hs2 \rightarrow fc_hs3$" and "$fc_hs4 \rightarrow fc_hs5 \rightarrow fc_hs6$" are the hash-code learning models $D_{img}$ and $D_{txt}$ of image and text modality, respectively. Pairwise Loss, quantization loss, and triple-supervised loss are objective functions.

## IV. THE PROPOSED APPROACH TDSRDH

In this section, we will demonstrate the details of TDSRDH, including model composition, learning optimization process, and algorithm details, etc.

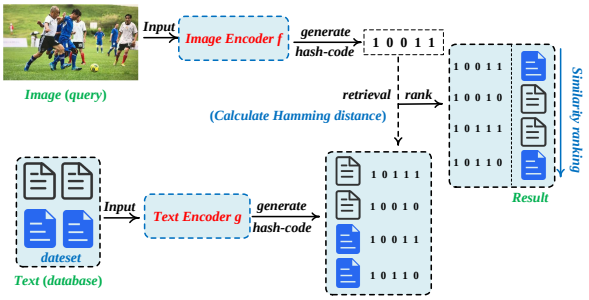



**FIGURE 3.** $Image \rightarrow Text$. Using the query of image modality retrieve similar items in the text database.

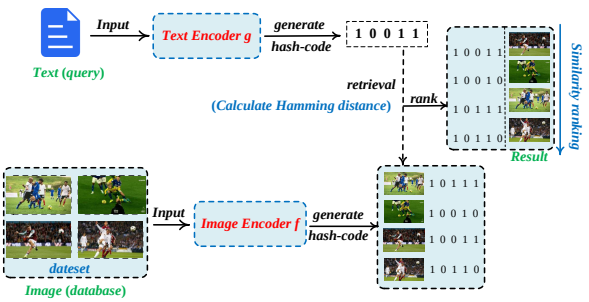



**FIGURE 4.** $Text \rightarrow Image$. Using the query of text modality retrieve similar items in the image database.

### A. MODEL

The whole TDSRDH model is illustrated in Fig. 2, which is an end-to-end learning framework that integrates feature learning and hash-code learning. The model mainly includes two parts: *Image Encoder* $f$ and *Text Encoder* $g$, which are introduced in detail in Section IV-A1. The hash-codes that are converted from TDSRDH are used in the retrieval process. The process of using the image to retrieve similar items in the text database ($Image \rightarrow Text$) is shown in Fig. 3. Similarly, the process of using the text to retrieve similar items in the image database ($Text \rightarrow Image$) is shown in Fig. 4. The retrieved results are ranked by Hamming distance.

#### 1) Feature Learning Part

In the feature learning part of TDSRDH, we use two networks to extract image and text modal features, respectively.

**Image Modal Encoder $f$:** The goal of the image encoder $f$ is to learn a mapping function $\boldsymbol{B}^{(v)} = f(\boldsymbol{V})$, which projects the image input $\boldsymbol{V}$ into the joint Hamming embedding space. It is worth noting that, as explained in Fig. 1, the cross-modal hashing method consists of two main parts, *i.e.*, feature learning and hash code learning. Among them, the chosen model mainly acts on the feature learning part, *i.e.*, feature extraction of the samples. It is worth stating that the subsequent experiments use the same model for feature extraction in all comparison experiments in order to ensure the fairness of the experiments. Inspired by the work [50], considering the model parameters, experimental environment, model performance, and other factors, as well as due to ResNet-152 [51] has the characteristics of easy



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



training and excellent performance in image classification, we choose it to extract image features. We use the ResNet-152 network pre-trained on the ImageNet dataset [52] to initialize the first 151 layers of the image encoder and freeze them, which finally outputs the learned image features. And we change the original 152nd layer of ResNet-152 to 4096 dimensions, which is also the beginning of the hash-code learning part. And on this basis (the 152nd layer), two fully connected layers are added, and the last fully connected layer outputs the learned hash-codes.

**TABLE 3.** Configuration of the image encoder $f$.

| Number | Layer | Configuration |
|---|---|---|
| 1 to 151 | $E_{img}$ | See ResNet-152 [51] |
| 152 | $fc_hs1$ | 4096 |
| 153 | $fc_hs2$ | 4096 |
| 154 | $fc_hs3$ | $k$(the length of hash-code) |

Table 3 manifests the detailed configuration of the image encoder $f$. Specifically, the feature learning model $E_{img}$ is composed of the first 151 layers of ResNet-152. In addition, the hash-code learning model $D_{img}$ is composed of fully connected layers: "$fc_hs1 \rightarrow fc_hs2 \rightarrow fc_hs3$". The number of each fully connected layer, for instance, "4096" indicates the number of nodes in this layer, and it also indicates the dimensionality of the output in this layer. It should be noted that for the last layer (the 154th layer) of the image encoder $f$, the identity function is selected as its activation function.

**Text Modal Encoder $g$:** The purpose of the text encoder $g$ is to learn a mapping function $\boldsymbol{B}^{(w)} = g(\boldsymbol{W})$, which projects the text input $\boldsymbol{W}$ into the joint Hamming embedding space. The majority of the previous work is based on multi-layer fully connected layers or LSTM [25] to encode text data. Here, we utilize a transformer-based encoder for text modal feature extraction due to it is widely used in natural language processing tasks and has superior performance.

Given a sentence sequence $\boldsymbol{W}_{j*} = \{w_1, w_2, \cdots, w_{d_y}\}$ composed of words, we try to encode it in a detailed and significant way for cross-modal retrieval and make full use of the learning flexibility of Transformers [29] to encode every text data. In this work, we use 1-layer 512-dimensional Transformers with 4 attention heads. First, we use *Word2Vec* to convert the original text into a word vector representation, where the embedding dimension is 512. And embed the position information of the word in the first layer. The feature representation is the average value of the output of the last layer of the Transformer.

On this basis, we add a 1024-dimensional fully connected layer "$fc_hs4$", which is also the beginning of the hash-code learning part. Similarly, two fully connected layers are added, and the last fully connected layer outputs the learned hash-codes. The detailed configuration of the text modal encoder is demonstrated in Table 4, where the configuration implies

**TABLE 4.** Configuration of the text encoder $g$.

| Number | Layer | Configuration |
|---|---|---|
| 1 | $E_{txt}$ | 512 |
| 2 | $fc_hs4$ | 1024 |
| 3 | $fc_hs5$ | 8192 |
| 4 | $fc_hs6$ | $k$(the length of hash-code) |

the number of dimensions. The hash-code learning model $D_{txt}$ is "$fc_hs4 \rightarrow fc_hs5 \rightarrow fc_hs6$". Among them, the activation functions of the second and third layers are ReLU [53], and the last layer is the identity function. Fig. 5 manifests that the structure of text modal encoder $g$ described above includes the feature extraction model $E_{txt}$ and the hash-code learning model $D_{txt}$.

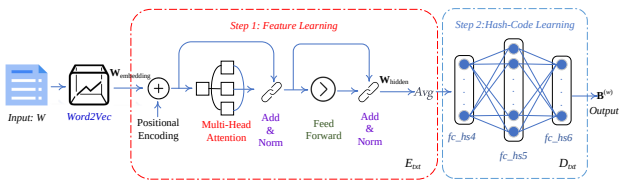



**FIGURE 5.** Structure of text encoder $g$. $E_{txt}$ is the feature learning model and $D_{txt}$ is the hash learning model.

### 2) Hash-Code Learning Part

Let $f(\boldsymbol{V}_{i*}; \theta_v, \theta_{v_hash}) \in \mathbb{R}^{1\times k}$ presents the hash-code learned by the image $\boldsymbol{V}_{i*}$, which is also the output of the image modal encoder $f$. Here, $\theta_v$ represents the parameters of $E_{img}$, and $\theta_{v_hash}$ represents the parameters of $D_{img}$. Similarly, let $g(\boldsymbol{W}_{j*}; \theta_w, \theta_{w_hash}) \in \mathbb{R}^{1\times k}$ presents the hash-code learned by the text $\boldsymbol{W}_{j*}$, which is also the output of the text modal encoder $g$. Here, $\theta_w$ represents the parameters of $E_{txt}$, and $\theta_{w_hash}$ represents the parameters of $D_{txt}$.

**Pairwise Loss:** We utilize the pairwise loss function to measure the similarity between hash-codes, and the cross-modal similarity in $\boldsymbol{S}$ can be preserved by optimizing this term [54], [55]. It is shared by the image modal encoder $f$ and the text modal encoder $g$, which can be designed as:

$$\min_{\theta_v,\theta_w,\theta_{v_hash},\theta_{w_hash}} \mathcal{L}_p = -\sum_{i,j=1}^{n} \left(\boldsymbol{S}_{ij}\Gamma_{ij} - \log\left(1 + e^{\Gamma_{ij}}\right)\right), \tag{3}$$

where $\Gamma_{ij} = \frac{1}{2}\boldsymbol{X}_{i*}\boldsymbol{Y}_{j*}^T$, $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ with $\boldsymbol{X}_{i*} = f(\boldsymbol{V}_{i*}; \theta_v, \theta_{v_hash})$, and $\boldsymbol{Y} \in \mathbb{R}^{n\times k}$ with $\boldsymbol{Y}_{j*} = g(\boldsymbol{W}_{j*}; \theta_w, \theta_{w_hash})$. Here, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the output of image and text hash-code learning models $D_{img}$ and $D_{txt}$, respectively. $\mathcal{L}_p$ is the negative log-likelihood, which can be derived by Eq.1 and Eq.2. It is also easy to obtain that minimizing this loss function is equal to maximizing likelihood [56], which makes the similarity (inner product) between $\boldsymbol{X}_{i*}$ and $\boldsymbol{Y}_{j*}$ the largest when $\boldsymbol{S}_{ij} = 1$, and the smallest when $\boldsymbol{S}_{ij} = 0$. Therefore, by optimizing this



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



item, the similarity between the $\boldsymbol{X}_{i*}$ and $\boldsymbol{Y}_{j*}$ in $\boldsymbol{S}$ can be preserved.

**Quantization Loss:** Simultaneously, to reduce the quantization error, we introduce the quantization loss, where the objective function of this part in the image encoder $f$ is designed as follows:

$$\begin{aligned} \min_{\theta_v,\theta_{v_hash}} \mathcal{L}_q^f &= \left\| \boldsymbol{B}^{(v)} - \boldsymbol{X} \right\|_F^2 = \|\boldsymbol{B} - \boldsymbol{X}\|_F^2, \\ \text{s.t.} \quad & \boldsymbol{B} \in \{-1, +1\}^{n \times k}, \end{aligned} \tag{4}$$

where $\boldsymbol{B}^{(v)} = sign(\boldsymbol{X})$. And the objective function of this part of the text encoder $g$ is designed as follows:

$$\begin{aligned} \min_{\theta_w,\theta_{w_hash}} \mathcal{L}_q^g &= \left\| \boldsymbol{B}^{(w)} - \boldsymbol{Y} \right\|_F^2 = \|\boldsymbol{B} - \boldsymbol{Y}\|_F^2, \\ \text{s.t.} \quad & \boldsymbol{B} \in \{-1, +1\}^{n \times k}, \end{aligned} \tag{5}$$

where $\boldsymbol{B}^{(w)} = sign(\boldsymbol{Y})$. Here, we consider that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the continuous surrogate of $\boldsymbol{B}^{(v)}$ and $\boldsymbol{B}^{(w)}$, respectively. Since $\boldsymbol{X}$ and $\boldsymbol{Y}$ can preserve cross-modal similarity in $\boldsymbol{S}$, the binary hash-codes $\boldsymbol{B}^{(v)}$ and $\boldsymbol{B}^{(w)}$ are also expected to preserve cross-modal similarity in $\boldsymbol{S}$, which is completely consistent with the principle of *preserving similarity* of cross-modal hashing. It is worth noting that, to ensure better performance during the training process, we set the hash-codes of the two modalities from the same training instance pair to be the same, $\boldsymbol{B}^{(v)} = \boldsymbol{B}^{(w)} = \boldsymbol{B}$. But after we have trained the model, we still need to generate different hash-codes $\boldsymbol{B}^{(v)}$ and $\boldsymbol{B}^{(w)}$ for two different modal data of query points or points from the database (not training points). It will be described in detail in Section IV-C.

**Triple-supervised Loss:** To distinguish between dissimilar samples as much as possible in the supervised learning process, inspired by AADAH [23], [57], [58], we propose a triple-supervised loss. The purpose is to make the distance between similar samples in the space as close as possible when the samples are mapped to the joint space, and the distance between different samples is as far as possible [59], which is illustrated in Fig. 6. It makes it possible to generate a discriminative representation as much as possible while preserving the similarity. The objective function of this part of the image modal encoder $f$ is designed as follows:

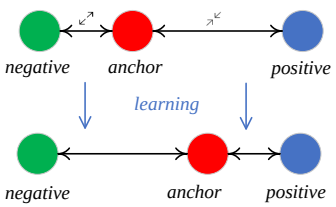



**FIGURE 6.** The variation of the distance between instances is under the influence of the triple-supervised loss.

$$\min_{\theta_v,\theta_{v_hash}} \mathcal{L}_s^f = \sum_{i,j=1}^{n} \mathcal{L}_t^f(i,j)\,\zeta(i,j), \tag{6}$$

where $\zeta(i,j) = 1$ if $i = j$, otherwise 0, and the $\mathcal{L}_t^f$ is defined as follows:

$$\mathcal{L}_t^f(i,j) = \max\left\{0, \varepsilon + \|\boldsymbol{X}_{i*} - \boldsymbol{Y}_{i*}\|_2^2 - \|\boldsymbol{X}_{i*} - \boldsymbol{Y}_{j*}\|_2^2\right\}, \tag{7}$$

where $\boldsymbol{X}_{i*}$ is the anchor sample, $\boldsymbol{Y}_{i*}$ is the positive sample, $\boldsymbol{Y}_{j*}$ is the negative sample, $\langle \boldsymbol{X}_{i*}, \boldsymbol{Y}_{i*}, \boldsymbol{Y}_{j*} \rangle$ is the triplet and the $\varepsilon$ is the margin. And the objective function of this part of the text modal encoder $g$ is designed as follows:

$$\min_{\theta_w,\theta_{w_hash}} \mathcal{L}_s^g = \sum_{i,j=1}^{n} \mathcal{L}_t^g(i,j)\,\zeta(i,j), \tag{8}$$

where the $\mathcal{L}_t^g$ is defined as follows:

$$\mathcal{L}_t^g(i,j) = \max\left\{0, \varepsilon + \|\boldsymbol{Y}_{j*} - \boldsymbol{X}_{j*}\|_2^2 - \|\boldsymbol{Y}_{j*} - \boldsymbol{X}_{i*}\|_2^2\right\}, \tag{9}$$

where $\boldsymbol{Y}_{j*}$ is the anchor sample, $\boldsymbol{X}_{j*}$ is the positive sample, $\boldsymbol{Y}_{i*}$ is the negative sample, $\langle \boldsymbol{Y}_{j*}, \boldsymbol{X}_{j*}, \boldsymbol{X}_{i*} \rangle$ is the triplet and the $\varepsilon$ is the margin. It is worth noting that all margins in this work are empirically set to 0.3.

It is worth noting that pairwise loss is the main objective function used by most of the existing works in preserving cross-modal similarity. There are also some novel works employing some customized triplet losses for preserving cross-modal similarity. Unlike other works, the triplet loss proposed in this work is based on the pairwise loss for preserving cross-modal similarity, the purpose is to map the sample to the common Hamming space, shortening the distance between similar samples and increasing the distance between dissimilar samples. In other words, the effect is to make the generated hash codes more discriminative, *i.e.*, the hash-code of similar samples are more similar and those of dissimilar samples are more dissimilar, which will significantly reduce retrieval errors to improve retrieval accuracy.

Therefore, combing Eq.3, Eq.4 and Eq.6, the objective function of the image modal encoder $f$ of TDSRDH can be defined as:

$$\min_{\theta_v,\theta_{v_hash}} \mathcal{L}^f = \mathcal{L}_p + \alpha\mathcal{L}_q^f + \beta\mathcal{L}_s^f. \tag{10}$$

Similarly, combing Eq.3, Eq.5 and Eq.8, the objective function of the text modal encoder $g$ can be defined as:

$$\min_{\theta_w,\theta_{w_hash}} \mathcal{L}^g = \mathcal{L}_p + \alpha\mathcal{L}_q^g + \beta\mathcal{L}_s^g. \tag{11}$$

where $\alpha$ and $\beta$ is a hyper-parameter. The overall objective function, combining the $\mathcal{L}^f$ and $\mathcal{L}^g$, is written as below:

$$\begin{aligned} \min_{\theta_v,\theta_w,\theta_{v_hash},\theta_{w_hash}} & \mathcal{L} = \mathcal{L}^f + \mathcal{L}^g, \\ \text{s.t.} \quad & \boldsymbol{B} \in \{-1, +1\}^{k \times n}. \end{aligned} \tag{12}$$

And, from Eq.12, we find that the parameters ($\theta_v$ and $\theta_w$) of ($E_{img}$ and $E_{txt}$) and the parameters ($\theta_{v_hash}$ and $\theta_{w_hash}$) of ($D_{img}$ and $D_{txt}$) are learned from the same objective function. This also indicates that TDSRDH is a deep learning framework that integrates feature learning and hash-code learning.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



### B. LEARNING

We employ an alternate-learning strategy to optimize the objective function in Eq.12, and learn $(\theta_v, \theta_{v_hash})$, $(\theta_w, \theta_{w_hash})$, and $\boldsymbol{B}$. Optimize the parameters of one encoder each time and fix the parameters of other encoders. Algorithm 1 summarizes the entire alternate learning optimization process of TDSRDH. In this subsection, we will introduce the detailed derivation process.

**Algorithm 1:** The alternate learning algorithm of TDSRDH

**Input:** image modal dataset $\boldsymbol{V}$, text modal dataset $\boldsymbol{W}$, multi-label matrix $\boldsymbol{L}$, cross-modal similarity matrix $\boldsymbol{S}$ and the length of hash-code $k$, parameters $\alpha$ and $\beta$

**Output:** image encoder $f$ parameters ($\theta_v$, $\theta_{v_hash}$), text encoder $g$ parameters ($\theta_w$, $\theta_{w_hash}$) and hash-code matrix $\boldsymbol{B}$

**Initialization:** encoder parameters $\theta_v$, $\theta_{v_hash}$, $\theta_w$ and $\theta_{w_hash}$, mini-batch size $N^{(v)} = N^{(w)} = 128$, the numer of iteration $t^{(v)} = \lceil \frac{n}{N^{(v)}} \rceil$, $t^{(w)} = \lceil \frac{n}{N^{(w)}} \rceil$, and the number of iteration epochs $n_e = 200$.

**while** *the cycle epoch iterates* $n_e$ *times* **do**

  **for** $iter = 1, 2, \cdots, t^{(v)}$ **do**

    • Randomly select $N^{(v)}$ instances from $\boldsymbol{V}$ to form a mini-batch;

    • For each sampling point $\boldsymbol{V}_{i*}$ (from the mini-batch), calculate $\boldsymbol{B}_{i*}^{(v)} = f(\boldsymbol{V}_{i*}; \theta_v, \theta_{v_hash})$ by forward-propagation;

    • Calculate the gradient according to Eq.20;

    • Use gradient back-propagation to update parameters $\theta_v$ and $\theta_{v_hash}$;

  **end for**

  **for** $iter = 1, 2, \cdots, t^{(w)}$ **do**

    • Randomly select $N^{(w)}$ instances from $\boldsymbol{W}$ to form a mini-batch;

    • For each sampling point $\boldsymbol{W}_{j*}$ (from the mini-batch), calculate $\boldsymbol{B}_{j*}^{(w)} = g(\boldsymbol{W}_{j*}; \theta_w, \theta_{w_hash})$ by forward-propagation;

    • Calculate the gradient according to Eq.21;

    • Use gradient back-propagation to update parameters $\theta_w$ and $\theta_{w_hash}$;

  **end for**

  Learn B according to Eq.13.

**end while**

#### 1) Learn $B$, with the parameters of $f$ and $g$ fixed

When $f$ and $g$ are fixed, the objective function in Eq.12 can be redefined as:

$$\max_{\boldsymbol{B}} \operatorname{tr}\left(\boldsymbol{B}^T(\alpha(\boldsymbol{X}+\boldsymbol{Y}))\right) = \operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{H}\right) = \sum_{i,j}\boldsymbol{B}_{ij}\boldsymbol{H}_{ij}, \quad \text{s.t.} \quad \boldsymbol{B} \in \{-1,+1\}^{k\times n}, \tag{13}$$

where $\boldsymbol{H} = \alpha(\boldsymbol{X}+\boldsymbol{Y})$. Here, we can find that if Eq.13 is to be guaranteed, the hash-code $\boldsymbol{B}_{ij}$ should maintain the same sign as $\boldsymbol{H}_{ij}$. Accordingly, $\boldsymbol{B}$ can be acquired by the following formula:

$$\boldsymbol{B} = \operatorname{sign}(\boldsymbol{H}) = \operatorname{sign}(\alpha(\boldsymbol{X}+\boldsymbol{Y})). \tag{14}$$

**Proof of the Eq.13:** The optimization of Eq.13 is derived from the optimization process of Eq.12 when both the image encoder *f* and the text encoder *g* are fixed. Specifically, when *f* and *g* are both fixed, we can remove coefficients and fixed terms that are not related to $\boldsymbol{B}^{(v)}$ and $\boldsymbol{B}^{(w)}$, and their changes will not affect the optimization of Eq.12. Accordingly, the optimization of Eq.13 at this time leaves the following equation:

$$\begin{aligned}\min_{\boldsymbol{B}} \mathcal{L} &= \min_{\boldsymbol{B}} \left(\mathcal{L}_q^f + \mathcal{L}_q^g\right) \\ &= \min_{\boldsymbol{B}} \left(\left\|\boldsymbol{B}^{(v)} - \boldsymbol{X}\right\|_F^2 + \left\|\boldsymbol{B}^{(w)} - \boldsymbol{Y}\right\|_F^2\right),\end{aligned} \tag{15}$$

It is mainly derived from Eq.4 and Eq.5.

Furthermore, as mentioned before, during training we set $\boldsymbol{B}^{(v)} = (\boldsymbol{w}) = (\boldsymbol{B})$. Then Eq.15 can be further simplified and deduced as:

$$\min_{\boldsymbol{B}} \mathcal{L} = \min_{\boldsymbol{B}} \left(\|\boldsymbol{B}-\boldsymbol{X}\|_F^2 + \|\boldsymbol{B}-\boldsymbol{Y}\|_F^2\right). \tag{16}$$

Moreover, taking the first term as an example, the Frobenius norm can be simplified by:

$$\begin{aligned}\|\boldsymbol{B}-\boldsymbol{X}\|_F^2 &\overset{def}{=} tr\left((\boldsymbol{B}-\boldsymbol{X})^T(\boldsymbol{B}-\boldsymbol{X})\right) \\ &= tr\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{B} - 2\boldsymbol{X}\boldsymbol{B}^T\right).\end{aligned} \tag{17}$$

Therefore, Eq.16 can be organized as:

$$\begin{aligned}\min_{\boldsymbol{B}} \mathcal{L} = \min_{\boldsymbol{B}} (&\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{B} - 2\boldsymbol{X}\boldsymbol{B}^T + \\ &\boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{B}^T\boldsymbol{B} - 2\boldsymbol{Y}\boldsymbol{B}^T).\end{aligned} \tag{18}$$

Since the elements in $\boldsymbol{B}$ are all 1 or -1, $\boldsymbol{B}^T\boldsymbol{B} = n \times c$. Also, $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be treated as constants since they are fixed terms. Further Eq.18 will be organized as:

$$\begin{aligned}\min_{\boldsymbol{B}} \mathcal{L} &= \min_{\boldsymbol{B}} tr - \left(\boldsymbol{B}^T(\boldsymbol{X}+\boldsymbol{Y})\right) \\ &= \max_{\boldsymbol{B}} \left(\boldsymbol{B}^T(\boldsymbol{X}+\boldsymbol{Y})\right) \\ &= tr\left(\boldsymbol{B}^T\boldsymbol{H}\right) \\ &= \sum_{i,j}\boldsymbol{B}_{ij}\boldsymbol{H}_{ij}.\end{aligned} \tag{19}$$



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



2) Learn $\theta_v$, $\theta_{v_hash}$, with $\boldsymbol{B}$ and text encoder $g$ fixed

We use mini-batch stochastic gradient descent (SGD) to learn the parameters $\theta_v$ and $\theta_{v_hash}$ of the image encoder $f$ with the back-propagation(BP) algorithm. Specifically, for each sampling point $\boldsymbol{V}_{i*}$, we calculate the gradient as shown in Eq. 20:

$$\begin{aligned}\frac{\partial \mathcal{L}^f}{\partial \boldsymbol{X}_{i*}} =& \frac{\partial \mathcal{L}_p}{\partial \boldsymbol{X}_{i*}} + \alpha \frac{\partial \mathcal{L}_q^f}{\partial \boldsymbol{X}_{i*}} + \beta \frac{\partial \mathcal{L}_s^f}{\partial \boldsymbol{X}_{i*}} \\ =& \frac{1}{2}\sum_{j=1}^{n} \left(\sigma\left(\Gamma_{ij}\right)\boldsymbol{Y}_{j*} - \boldsymbol{S}_{ij}\boldsymbol{Y}_{j*}\right) + 2\alpha(\boldsymbol{X}_{i*} - \boldsymbol{B}_{i*}) \\ &+ 2\beta\sum_{j=1}^{n}(\boldsymbol{Y}_{j*} - \boldsymbol{Y}_{i*}).\end{aligned} \tag{20}$$

Then we use the chain rule to calculate $\frac{\partial \mathcal{L}}{\partial \theta_v}$ and $\frac{\partial \mathcal{L}}{\partial \theta_{v_hash}}$ with $\frac{\partial \mathcal{L}}{\partial \boldsymbol{X}_{i*}}$, and update the parameters $\theta_v$ and $\theta_{v_hash}$ through BP.

3) Learn $\theta_w$, $\theta_{w_hash}$, with $\boldsymbol{B}$ and image encoder $f$ fixed

We also use mini-batch SGD to learn the parameters $\theta_w$ and $\theta_{w_hash}$ of the text encoder $g$ with the BP algorithm. Specifically, for each sampling point $\boldsymbol{W}_{j*}$, we calculate the gradient as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}^g}{\partial \boldsymbol{Y}_{j*}} =& \frac{\partial \mathcal{L}_p}{\partial \boldsymbol{Y}_{j*}} + \alpha \frac{\partial \mathcal{L}_q^g}{\partial \boldsymbol{Y}_{j*}} + \beta \frac{\partial \mathcal{L}_s^g}{\partial \boldsymbol{Y}_{j*}} \\ =& \frac{1}{2}\sum_{i=1}^{n} \left(\sigma\left(\Gamma_{ij}\right)\boldsymbol{X}_{i*} - \boldsymbol{S}_{ij}\boldsymbol{X}_{i*}\right) + 2\alpha(\boldsymbol{Y}_{j*} - \boldsymbol{B}_{j*}) \\ &+ 2\beta\sum_{i=1}^{n}(\boldsymbol{X}_{i*} - \boldsymbol{X}_{j*}).\end{aligned} \tag{21}$$

Then we use the chain rule to calculate $\frac{\partial \mathcal{L}}{\partial \theta_w}$ and $\frac{\partial \mathcal{L}}{\partial \theta_{w_hash}}$ with $\frac{\partial \mathcal{L}}{\partial \boldsymbol{Y}_{j*}}$, and update the parameters $\theta_w$ and $\theta_{w_hash}$ through BP.

### C. OUT-OF-SAMPLE EXTENSION

For any data not in the training dataset, we generate different hash-codes according to their different modal types. Specifically, given an image data point (not training data) $\boldsymbol{v}_i$, the hash-code is generated through the forward propagation of the image encoder $f$ as:

$$\boldsymbol{b}_i^{(v)} = sign(f(\boldsymbol{v}_i; \theta_v, \theta_{v_hash})). \tag{22}$$

Similarly, if a text data point (not training data) is given $\boldsymbol{w}_j$, the hash-code is generated through the forward propagation of the text encoder $g$ as:

$$\boldsymbol{b}_j^{(w)} = sign(g(\boldsymbol{w}_j; \theta_w, \theta_{w_hash})). \tag{23}$$

## V. EXPERIMENT

In this section, we conduct our experiments on three widely used public datasets and compare them with some typical methods.

### A. DATASETS

The three datasets used for evaluation are *MIRFLICKR-25K* [60], *IAPR TC-12* [61], and *NUS-WIDE* [62], respectively.

The *MIRFLICKR-25K* dataset [60] contains 25,000 image-text pairs from the Flickr website. Here, we follow the setting in the DCMH [22] experiment and select 20,015 image-text pairs containing at least 20 textual tags. Among them, each data point is annotated with at least 24 unique labels.

The *IAPR TC-12* dataset [61] includes 20,000 image-text pairs, each of which uses 255 labels. In our experiment, we use the entire dataset.

The *NUS-WIDE* dataset [62] contains 269,648 web images associated with textual tags, and each data point has one or more of 81 unique labels. We also follow the setting in DCMH [22] and select 195,834 image-text pairs that belong to the 21 most frequent concepts.

It is worth noting that for all datasets, if the image data point $\boldsymbol{V}_i$ and the text data point $\boldsymbol{W}_j$ share at least one common label, then $\boldsymbol{V}_i$ and $\boldsymbol{W}_j$ are considered similar, and $\boldsymbol{S}_{ij} = 1$. Otherwise, it is considered that $\boldsymbol{V}_i$ and $\boldsymbol{W}_j$ are not similar, $\boldsymbol{S}_{ij} = 0$.

### B. EVALUATION METRIC

1) Evaluation Protocol

For the *MIRFLICKR-25K* and *IAPR TC-12* dataset, randomly extract 2,000 instances (image-text pairs) serve as the test set (query), the remaining data are used as the retrieval set (database), and finally 10,000 pairs of examples are randomly extracted from the retrieval set as the training set. For the *NUS-WIDE* dataset, 2,100 instances are randomly extracted as the test set, the remaining are used as the retrieval set, and finally 10,500 pairs of instances are randomly extracted from the retrieval set as the training set. The statistics on the division of each dataset are shown in Table 5.

**TABLE 5.** Dataset division

| **Dadaset** | *MIRFLICKR-25K* | *IAPR TC-12* | *NUS-WIDE* |
|---|---|---|---|
| training | 10,000 | 10,000 | 10,500 |
| test | 2,000 | 2,000 | 2,100 |
| retrieval | 18,015 | 18,000 | 193,734 |
| total | 20,015 | 20,000 | 195,834 |

In order to evaluate the performance of TDSRDH, we employ two widely used retrieval protocols, *Hamming ranking* and *hash lookup* [17]. Among them, the former sorts the Hamming distances to the points in the database (retrieval) in ascending order according to a given query point. Mean average precision (MAP) is the mean value of average precision (AP) and is used to measure the *Hamming ranking* protocol. The definition of AP value is as follows:

$$\mathrm{AP} = \frac{1}{k}\sum_{i=1}^{k}\frac{r_i}{i}, \tag{24}$$



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



where $k$ denotes the number of instances related to the current test in the retrieval set, and $r_i$ indicates the number of instances in the top $i$ related results. Therefore, the definition of MAP is as follows:

$$\mathrm{MAP} = \frac{1}{q}\sum_{j=1}^{q} \mathrm{AP}_j, \tag{25}$$

where $q$ denotes the number of query set, and $AP_j$ represents the AP value of the *j*-th query instance.

Through the *hash lookup* protocol, all the points within a given Hamming radius from the query point can be obtained. We employ the precision-recall (PR) curve to measure the performance of the hash lookup protocol, which is a widely used evaluation metric.

#### 2) Baseline

TDSRDH is compared with eight state-of-the-art methods, including the methods based on shallow structure(CCA [4], SCM [15], SePH [16]) and the methods based on deep structure(DCMH [22], AADAH [23], RDCMH [39], MGAH [42], CPAH [43], URGCH [7]). For a fair comparison, the deep features extracted by the pre-trained ResNet-152 are used as the image input for all baselines based on shallow-structure. Except for MGAH and CPAH, the source codes of all baselines are derived from the open-source code kindly provided by the original author. For MGAN and CPAH, since the code of these methods is not yet publicly available, we implement it carefully by ourselves according to the author's suggestion in the original paper [7], [42], [43]. In order to ensure the rigor and persuasiveness of the results, all baselines parameters are set according to the open-source code provided by the original author or the suggestions in the original paper, and all comparison baselines use the same pre-processed dataset after the same division. For all baselines, we try our best to eliminate all interference that affects the experiment, ensure that all baselines run in the same environment, and compare the results with scientific rigor. In SePH, according to the author's suggestion, RBF kernel are used and 500 points are randomly selected as core bases. In CPAH, image feature extraction includes two CNN architectures, we choose ResNet-152 architecture to extract image representation. In RDCMH, we do not mask the labels of the training data, and use all the labels for training.

### C. RELATED PARAMETERS SETTING

For TDSRDH, We use the ResNet-152 network [51] pre-trained on the ImageNet dataset [52] to initialize the image encoder $f$. In addition, all other parameters of the network in TDSRDH are initialized randomly. The input of the image modal encoder $f$ uses $224 \times 224$ primitive pixels, and the text modal encoder $g$ uses original text sentence input to *Word2Vec* to obtain the 512-dimensional word embedding vectors. Then, the number of iterations is set to 200, the learning rate is selected from $0.00001$ to $0.1$ with the validation set. Similarly, we use the validation set to select hyperparameters $\alpha$ and $\beta$. Moreover, the batch size is fixed to 128. In addition, to eliminate the randomness of the experiment and maintain rigor, for TDSRDH, we conduct ten experiments to average the results. Finally, all experiments are run in the GPU environment of Tesla P100.

**TABLE 6.** MAPs of Hamming ranking on the *MIRFLICKR-25K*. The best MAPs are shown in boldface and the second best MAPs are underlined.

| Task | Method | *MIRFLICKR-25K* | | |
|---|---|---|---|---|
| | | 16 bits | 32 bits | 64 bits |
| $Image \rightarrow Text$ | **CCA** | 0.5707 | 0.5688 | 0.5663 |
| | **SCM** | 0.6841 | 0.6919 | 0.7000 |
| | **SePH** | 0.7112 | 0.7186 | 0.7230 |
| | **DCMH** | 0.7401 | 0.7457 | 0.7506 |
| | **AADAH** | 0.7567 | 0.7723 | 0.7728 |
| | **RDCMH** | 0.7722 | 0.7735 | 0.7786 |
| | **MGAN** | 0.6864 | 0.6932 | 0.7051 |
| | **CPAH** | **0.7736** | 0.7917 | 0.7873 |
| | **URGCH** | 0.7613 | <u>0.7922</u> | <u>0.7907</u> |
| | **TDSRDH** | <u>0.7719</u> | **0.7953** | **0.7937** |
| $Text \rightarrow Image$ | **CCA** | 0.5739 | 0.5705 | 0.5680 |
| | **SCM** | 0.6930 | 0.7003 | 0.7047 |
| | **SePH** | 0.7211 | 0.7256 | 0.7312 |
| | **DCMH** | 0.7831 | 0.7906 | 0.7940 |
| | **AADAH** | 0.7928 | <u>0.8070</u> | 0.8082 |
| | **RDCMH** | <u>0.7929</u> | 0.7924 | 0.8001 |
| | **MGAN** | 0.6743 | 0.6787 | 0.6985 |
| | **CPAH** | 0.7775 | 0.7870 | 0.7901 |
| | **URGCH** | 0.7849 | 0.7950 | <u>0.8153</u> |
| | **TDSRDH** | **0.8043** | **0.8125** | **0.8222** |

### D. PERFORMANCE

#### 1) Hamming Ranking

Table 6, 7 and 8 report the experimental MAP results of TDSRDH and other baselines on three datasets(*MIRFLICKR-25K*, *IAPR TC-12* and *NUS-WIDE*), respectively. Here, "$Image \rightarrow Text$" means that the query set is image modality and the retrieval set is text modality. Similarly, "$Text \rightarrow Image$" means that the query set is text modality and the retrieval set is image retrieval. From the results in the Tables, we can find that the performance of deep structure-based methods is comprehensive and significantly better than that of shallow structure-based methods, which illustrates the superiority of deep structure-based methods. Furthermore, TDSRDH can achieve better performance than other baselines. For example, on IAPR TC-12, the MAP of our method is 0.5742, compared to the value of 0.5582 for the second-best method (URGCH), on 64 bits when the image is used as the query to retrieve text. Since in the training process, TDSRDH can generate hash-codes that are more discriminative and preserve similarity. In addition, the performance of the model is further improved through the



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



**TABLE 7.** MAPs of Hamming ranking on the *IAPR TC-12*. The best MAPs are shown in boldface and the second best MAPs are underlined.

| Task | Method | *IAPR TC-12* | | |
|---|---|---|---|---|
| | | 16 bits | 32 bits | 64 bits |
| $Image \rightarrow Text$ | **CCA** | 0.3417 | 0.3353 | 0.3296 |
| | **SCM** | 0.3688 | 0.3659 | 0.3800 |
| | **SePH** | 0.4434 | 0.4546 | 0.4633 |
| | **DCMH** | 0.4521 | 0.4725 | 0.4839 |
| | **AADAH** | 0.5301 | 0.5291 | 0.5447 |
| | **RDCMH** | 0.5248 | 0.5287 | 0.5392 |
| | **MGAN** | 0.5263 | 0.5307 | 0.5400 |
| | **CPAH** | 0.5375 | 0.5387 | 0.5534 |
| | **URGCH** | 0.5398 | 0.5443 | 0.5582 |
| | **TDSRDH** | **0.5492** | **0.5464** | **0.5742** |
| $Text \rightarrow Image$ | **CCA** | 0.3488 | 0.3433 | 0.3374 |
| | **SCM** | 0.3451 | 0.3406 | 0.3467 |
| | **SePH** | 0.4419 | 0.4525 | 0.4643 |
| | **DCMH** | 0.5183 | 0.5377 | 0.5470 |
| | **AADAH** | 0.5360 | 0.5572 | 0.5654 |
| | **RDCMH** | 0.5327 | 0.5529 | 0.5630 |
| | **MGAN** | 0.5430 | 0.5564 | 0.5647 |
| | **CPAH** | 0.5506 | 0.5670 | 0.5684 |
| | **URGCH** | 0.5487 | 0.5591 | 0.5693 |
| | **TDSRDH** | **0.5519** | **0.5783** | **0.5755** |

**TABLE 8.** MAPs of Hamming ranking on the *NUS-WIDE*. The best MAPs are shown in boldface and the second best MAPs are underlined.

| Task | Method | *NUS-WIDE* | | |
|---|---|---|---|---|
| | | 16 bits | 32 bits | 64 bits |
| $Image \rightarrow Text$ | **CCA** | 0.3597 | 0.3475 | 0.3383 |
| | **SCM** | 0.5486 | 0.5480 | 0.5549 |
| | **SePH** | 0.6038 | 0.6134 | 0.6206 |
| | **DCMH** | 0.5900 | 0.6028 | 0.6091 |
| | **AADAH** | 0.6410 | 0.6297 | 0.6523 |
| | **RDCMH** | 0.6231 | 0.6236 | 0.6273 |
| | **MGAN** | 0.6132 | 0.6287 | 0.6310 |
| | **CPAH** | 0.6264 | 0.6637 | 0.6719 |
| | **URGCH** | 0.6385 | 0.6519 | 0.6724 |
| | **TDSRDH** | **0.6662** | **0.6681** | **0.6770** |
| $Text \rightarrow Image$ | **CCA** | 0.3611 | 0.3490 | 0.3387 |
| | **SCM** | 0.5341 | 0.5408 | 0.5481 |
| | **SePH** | 0.5976 | 0.6017 | 0.6102 |
| | **DCMH** | 0.6380 | 0.6505 | 0.6563 |
| | **AADAH** | 0.6792 | 0.6977 | 0.7041 |
| | **RDCMH** | 0.6641 | 0.6685 | 0.6694 |
| | **MGAN** | 0.6148 | 0.6402 | 0.6411 |
| | **CPAH** | 0.6146 | 0.6490 | 0.6615 |
| | **URGCH** | 0.6536 | 0.6824 | 0.7083 |
| | **TDSRDH** | **0.6995** | **0.7016** | **0.7147** |

**TABLE 9.** Ablation experiment of text encoder $g$ and triple-supervised loss $\mathcal{L}_s$ on *MIRFLICKR-25K*.

| Task | Method | *MIRFLICKR-25K* | | |
|---|---|---|---|---|
| $I \rightarrow T$ | *Ablation-I* | 0.7440 | 0.7604 | 0.7636 |
| | *Ablation-II* | 0.7626 | 0.7832 | 0.7853 |
| | **TDSRDH** | 0.7719 | 0.7953 | 0.7937 |
| $T \rightarrow I$ | *Ablation-I* | 0.7840 | 0.7918 | 0.7955 |
| | *Ablation-II* | 0.7990 | 0.8039 | 0.8130 |
| | **TDSRDH** | 0.8043 | 0.8125 | 0.8222 |

strong representation of images and text, so it can mine the associations between different modal data more effectively.

#### 2) Hash lookup

In this protocol, given any hash radius, the values of precision and recall can be calculated. By changing the radius from 0 to $k$ in step 1, we can obtain the PR curve. As shown in Fig. 7, it records a PR curve with a hash-code length of 16. The higher the position of the PR curve, the better the retrieval performance. Therefore, on three widely used datasets, TDSRDH outperforms other baselines.

### E. HYPERPARAMETER ANALYSIS

We also conducted experiments to explore the influence of hyperparameters $\alpha$ and $\beta$ on the results. Fig. 8 illustrates the MAP value results of $\alpha$ and $\beta$ with different values on *MIRFLICKR-25K* dataset, where the hash-code length is 16. According to the experimental results in the validation set, we set $\alpha = \beta = 1$ in the experiment.

### F. ABLATION EXPERIMENT

To verify the effectiveness of our proposed text encoder $g$ and the loss function in the model (for triple-supervised loss $\mathcal{L}_s$), we conduct ablation experiments on the *MIRFLICKR-25K* dataset and *IAPR TC-12* dataset. The results are manifested in Table 9 and Table 10. "$Ablation\text{-}I$" means that the text model merely uses the double-layer fully connected layer with the same settings as in DCMH [22], and does not use the tripe-supervised loss $\mathcal{L}_s$ ($\beta$ is set to 0). "$Ablation\text{-}II$" means that we only used the proposed text encoder $g$. Compared with "$Ablation\text{-}I$", it brings about $2\%$ improvement, which proves the effectiveness of our proposed text encoder $g$. "TDSRDH" means that we use both the text encoder $g$ and the triple-supervised loss $\mathcal{L}_s$. Compared with "$Ablation\text{-}II$", it brings about $1\%$ improvement, which proves the effectiveness of our proposed triple-supervised loss $\mathcal{L}_s$. In summary, the TDSRDH we proposed is effective.

### G. CONVERGENCE ANALYSIS

In order to verify the convergence of the proposed TDSRDH, we conduct an experiment on the large-scale dataset *NUS-WIDE* with the hash-code length of 32. The experimental results are illustrated in Fig. 9, which shows in detail the



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



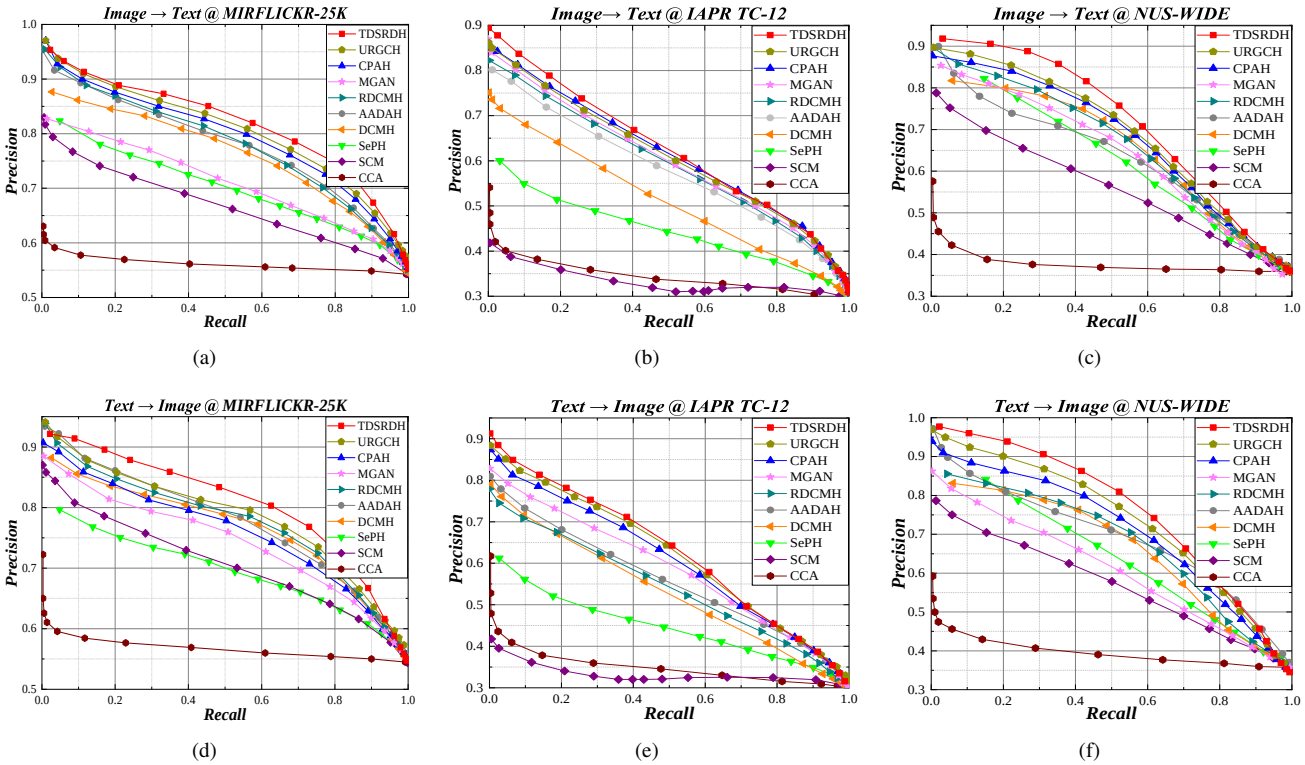



**FIGURE 7.** Precision-recall curves on three datasets(the hash-code length is 16).

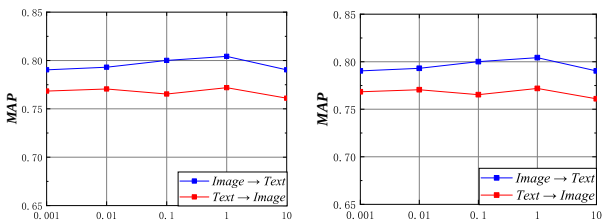



**FIGURE 8.** The influence of different values of hyperparameters $\alpha$ and $\beta$ on MAP, which is on the *MIRFLICKR-25K* dataset and the hash-code length is 16 bits.

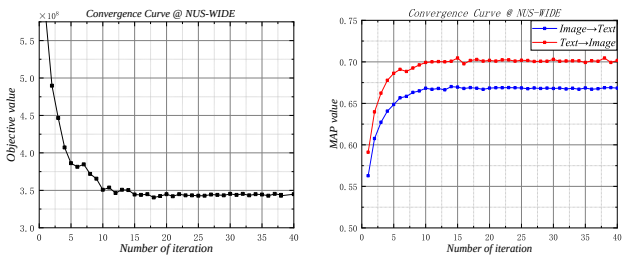



**FIGURE 9.** Results of objective function and MAP values of TDSRDH recorded in different iterations (on the NUS-WIDE with 32 bits).

**TABLE 10.** Ablation experiment of text encoder $g$ and triple-supervised loss $\mathcal{L}_s$ on *IAPR TC-12*.

| Task | Method | *IAPR TC-12* | | |
|---|---|---|---|---|
| $I \rightarrow T$ | *Ablation-I* | 0.5187 | 0.5106 | 0.5382 |
| | *Ablation-II* | 0.5403 | 0.5337 | 0.5619 |
| | **TDSRDH** | 0.5492 | 0.5464 | 0.5742 |
| $T \rightarrow I$ | *Ablation-I* | 0.5264 | 0.5503 | 0.5480 |
| | *Ablation-II* | 0.5430 | 0.5689 | 0.5632 |
| | **TDSRDH** | 0.5519 | 0.5783 | 0.5755 |

convergence of the objective function and the MAP values. In Fig. 9(a), it can be found that the objective function values can converge quickly, at about 10 iterations. In Fig. 9(b), "$Image \rightarrow Text$" means to retrieve text through image queries, and "$Text \rightarrow image$" means to retrieve image through text queries, and it can be found that the MAP values of the two retrieval tasks can also converge quickly. This signifies that in the training process, the two MAP values increase as the objective function values decrease and finally converge. Therefore, with the rapid convergence of the proposed TDSRDH, the computational load of the model can also be significantly reduced.

### H. TRAINING TIME

The computational complexity mainly comes from the training process of the model. Therefore, to evaluate the complexity of the proposed TDSRDH, we record the training time spent on TDSRDH and each baseline on the *MIRFLICKR-25K* dataset and finally compare to further verify the com-



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



**TABLE 11.** Training time (in second) based on deep-structure baselines and TDSRDH on *MIRFLICKR-25K* dataset.

| Method | 16 bits | 32 bits | 64 bits |
|---|---|---|---|
| **DCMH** | 73.3527 | 73.8730 | 73.8713 |
| **AADAH** | 78.1539 | 78.9435 | 79.0561 |
| **RDCMH** | 73.5028 | 74.3816 | 75.7820 |
| **MGAN** | 76.8544 | 77.1879 | 77.1904 |
| **CPAH** | 75.3011 | 75.3450 | 75.3504 |
| **TDSRDH** | 74.0450 | 74.1539 | 74.2010 |

plexity. It is worth noting that in the training process, the feature extraction part of the model based on the shallow-structure is separated from the hash-codes learning part. The difference is that the deep-structure integrates the feature extraction and the hash-codes learning process, which makes the methods based deep-structure usually more time-consuming than the shallow-structure model. And through the previous reasoning, it can be known that the performance of the baselines based on the shallow-structure is not as good as the deep-structure. Therefore, in this experiment, we only compare baselines based on deep-structure. Similarly, in order to ensure a fair comparison, we all use the same experimental settings as described above, and conduct experiments under the same GPU environment. Table 11 reports the results of TDSRDH compared to the baselines. From the table, we can observe that, compared with the advanced deep-structure baselines, although the training time of the proposed TDSRDH is not the least, it does not have much difference compared with the best baseline. Comfortingly, the performance improvement of TDSRDH is satisfactory. Therefore, compared with the existing cross-modal hashing methods, TDSRDH has more competitive computational efficiency and performance.

### I. INNOVATION COMPARISON

To provide further clarity on the innovations of this work, a table comparing the innovations with the deep structure-based baselines is shown in Table 12. Therefore, after the previous research and analysis, the main innovations can be summarised as: TDSRDH seamlessly integrates feature learning and hash-code learning, each end of which is based on the high-performance encoder to obtain a strong representation of the features, where we innovatively utilize a transformer-based encoder to encode the text modality. Moreover, we propose a triple-supervised loss and combine it with pairwise loss and quantization loss, so that TDSRDH can generate more discriminative representations while preserving the similarity.

## VI. CONCLUSION

In this paper, we have proposed a novel cross-modal hashing method, called Transformer-based Discriminative and Strong Representation Deep Hashing for Cross-Modal Retrieval (TDSRDH), which is based on the end-to-end model of high-performance text and image encoders. It can make full use of the semantic consistency and correlation between different modalities. Furthermore, integrating with pairwise loss, quantization loss, and triple-supervised loss can generate more discriminative representations while preserving the similarity between modal data. Specifically, it can make the distance between similar samples in the space as close as possible when there are mapped to the joint space, and the distance between different samples is as far as possible. Finally, we conduct extensive experiments on three widely used datasets(*MIRFLICKR-25K*, *IAPR TC-12*, and *NUS-WIDE*), and the results demonstrate that our proposed TDSRDH has achieved satisfactory performance.

## REFERENCES

[1] Y. Duan, N. Chen, S. Shen, P. Zhang, Y. Qu, and S. Yu, "Fdsa-stg: Fully dynamic self-attention spatio-temporal graph networks for intelligent traffic flow prediction," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2022.
[2] P. Zhang, N. Chen, S. Shen, S. Yu, N. Kumar, and C.-H. Hsu, "Ai-enabled space-air-ground integrated networks: Management and optimization," *IEEE Network*, pp. 1–7, 2023.
[3] K. Zhou, L. Zhao, Y. Zigao, H. Wang, T. Lin, F. Sheng, and Y. Yufen, "Equal value string and copy above string based string prediction for scc in avs3," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.
[4] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.
[5] Y. Zhang, C. Li, N. Chen, and P. Zhang, "Intelligent requests orchestration for microservice management based on blockchain in software defined networking: a security guarantee," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2022, pp. 254–259.
[6] N. Chen, S. Shen, Y. Duan, S. Huang, W. Zhang, and L. Tan, "Non-euclidean graph-convolution virtual network embedding for space–air–ground integrated networks," *Drones*, vol. 7, no. 3, p. 165, Feb 2023.
[7] Y. Duan, N. Chen, A. K. Bashir, M. D. Alshehri, L. Liu, P. Zhang, and K. Yu, "A web knowledge-driven multimodal retrieval method in computational social systems: Unsupervised and robust graph convolutional hashing," *IEEE Transactions on Computational Social Systems*, pp. 1–11, 2022.
[8] T. Li, H. Wang, D. He, and J. Yu, "Synchronized provable data possession based on blockchain for digital twin," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 472–485, 2022.
[9] P. Zhang, N. Chen, S. Shen, S. Yu, S. Wu, and N. Kumar, "Future quantum communications and networking: A review and vision," *IEEE Wireless Communications*, pp. 1–8, 2022.
[10] L. Zhao, T. Lin, D. Zhang, K. Zhou, and S. Wang, "An ultra-low complexity and high efficiency approach for lossless alpha channel coding," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 786–794, 2020.
[11] Y. Duan, N. Chen, P. Zhang, N. Kumar, L. Chang, and W. Wen, "Ms2gah: Multi-label semantic supervised graph attention hashing for robust cross-modal retrieval," *Pattern Recognition*, vol. 128, p. 108676, 2022.
[12] X. Shi, F. Xing, K. Xu, M. Sapkota, and L. Yang, "Asymmetric discrete graph hashing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
[13] 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*, 2014, pp. 2075–2082.
[14] L. Ma, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Global and local semantics-preserving based deep hashing for cross-modal retrieval," *Neurocomputing*, vol. 312, pp. 49–62, 2018.
[15] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1, 2014.
[16] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3864–3872.
[17] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," *Advances in neural information processing systems*, vol. 27, 2014.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



**TABLE 12.** Innovation comparison with deep structure-based baselines.

| Comparison | Image Encoder | Text Encoder | Objective Function |
|---|---|---|---|
| **DCMH** | CNN-F + fully-connected layer | BOW + fully-connected layer | Pairwise loss + Quantization loss |
| **AADAH** | VGGNet + attention mechanism | BOW + attention mechanism | Triplet ranking loss |
| **RDCMH** | CNN-F + fully-connected layer | BOW + fully-connected layer | Triplet ranking loss + Quantization loss |
| **MGAH** | Generative adversarial network | Generative adversarial network | Triplet ranking loss |
| **CPAH** | CNN + adversarial learning module | BOW + adversarial learning module | Pairwise loss + Quantization Loss + Classification loss |
| **URGCH** | Graph convolutional network + fully-connected layer | Graph convolutional network + fully-connected layer | Pairwise loss + Quantization loss |
| **TDSRDH** | Resnet-152 + fully-connected layer | Transformer + fully-connected layer | Pairwise loss + Triple-supervised loss + Quantization loss |

[18] N. Chen, P. Zhang, N. Kumar, C.-H. Hsu, L. Abualigah, and H. Zhu, "Spectral graph theory-based virtual network embedding for vehicular fog computing: A deep reinforcement learning architecture," *Knowledge-Based Systems*, vol. 257, p. 109931, 2022.

[19] P. Zhang, N. Chen, S. Li, K.-K. R. Choo, C. Jiang, and S. Wu, "Multi-domain virtual network embedding algorithm based on horizontal federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 3363–3375, 2023.

[20] P. Zhang, N. Chen, N. Kumar, L. Abualigah, M. Guizani, Y. Duan, J. Wang, and S. Wu, "Energy allocation for vehicle-to-grid settings: A low-cost proposal combining drl and vne," *IEEE Transactions on Sustainable Computing*, pp. 1–13, 2023.

[21] Y. Duan, N. Chen, L. Chang, Y. Ni, S. V. N. S. Kumar, and P. Zhang, "Capso: Chaos adaptive particle swarm optimization algorithm," *IEEE Access*, vol. 10, pp. 29 393–29 405, 2022.

[22] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3232–3240.

[23] X. Zhang, H. Lai, and J. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 591–606.

[24] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] H. Wang, D. Sahoo, C. Liu, K. Shu, P. Achananuparp, E.-p. Lim, and C. S. Hoi, "Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism," *IEEE Transactions on Multimedia*, 2021.

[27] Q. Zhou, L. Zhao, K. Zhou, T. Lin, H. Wang, S. Wang, and M. Jiao, "String prediction for 4:2:0 format screen content coding and its implementation in avs3," *IEEE Transactions on Multimedia*, vol. 23, pp. 3867–3876, 2021.

[28] P. Zhang, X. Huang, and L. Zhang, "Information mining and similarity computation for semi-un-structured sentences from the social data," *Digital Communications and Networks*, vol. PP, no. 99, pp. 1–1, 2021.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[30] A. Salvador, E. Gundogdu, L. Bazzani, and M. Donoser, "Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 475–15 484.

[31] L. Wang, Z. Zheng, N. Chen, Y. Chi, Y. Liu, H. Zhu, P. Zhang, and N. Kumar, "Multi-target-aware energy orchestration modeling for grid 2.0: A network virtualization approach," *IEEE Access*, vol. 11, pp. 21 699–21 711, 2023.

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

[33] Q. Li, L. Li, W. Wang, Q. Li, and J. Zhong, "A comprehensive exploration of semantic relation extraction via pre-trained cnns," *Knowledge-Based Systems*, vol. 194, p. 105488, 2020.

[34] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 964–981, 2021.

[35] M. Meng, H. Wang, J. Yu, H. Chen, and J. Wu, "Asymmetric supervised consistent and specific hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 986–1000, 2020.

[36] K. Zhan, N. Chen, S. V. N. Santhosh Kumar, G. Kibalya, P. Zhang, and H. Zhang, "Edge computing network resource allocation based on virtual network embedding," *International Journal of Communication Systems*, vol. n/a, no. n/a, p. e5344.

[37] Q. Lin, W. Cao, Z. He, and Z. He, "Semantic deep cross-modal hashing," *Neurocomputing*, vol. 396, pp. 113–122, 2020.

[38] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 394–10 403.

[39] X. Liu, G. Yu, C. Domeniconi, J. Wang, Y. Ren, and M. Guo, "Ranking-based deep cross-modal hashing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4400–4407.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[41] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4242–4251.

[42] 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.

[43] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 3626–3637, 2020.

[44] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4626–4634.

[45] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised contrastive cross-modal hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3877–3889, 2023.

[46] X. Liu, H. Zeng, Y. Shi, J. Zhu, C.-H. Hsia, and K.-K. Ma, "Deep cross-modal hashing based on semantic consistent ranking," *IEEE Transactions on Multimedia*, pp. 1–12, 2023.

[47] F. Yang, Y. Liu, X. Ding, F. Ma, and J. Cao, "Asymmetric cross–modal



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581



hashing with high–level semantic similarity," *Pattern Recognition*, vol. 130, p. 108823, 2022.

[48] S. Teng, C. Ning, W. Zhang, N. Wu, and Y. Zeng, "Fast asymmetric and discrete cross-modal hashing with semantic consistency," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 2, pp. 577–589, 2023.

[49] F. Wu, S. Li, G. Gao, Y. Ji, X.-Y. Jing, and Z. Wan, "Semi-supervised cross-modal hashing via modality-specific and cross-modal graph convolutional networks," *Pattern Recognition*, vol. 136, p. 109211, 2023.

[50] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[54] L. Ma, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Discriminative deep metric learning for asymmetric discrete hashing," *Neurocomputing*, vol. 380, pp. 115–124, 2020.

[55] L. Ma, X. Li, Y. Shi, L. Huang, Z. Huang, and J. Wu, "Learning discrete class-specific prototypes for deep semantic hashing," *Neurocomputing*, vol. 443, pp. 85–95, 2021.

[56] L. Ma, X. Li, Y. Shi, J. Wu, and Y. Zhang, "Correlation filtering-based hashing for fine-grained image retrieval," *IEEE Signal Processing Letters*, vol. 27, pp. 2129–2133, 2020.

[57] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[58] L. Ma, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Learning efficient binary codes from high-level feature representations for multilabel image retrieval," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2545–2560, 2017.

[59] L. Ma, H. Hong, F. Meng, Q. Wu, and J. Wu, "Deep progressive asymmetric quantization based on causal intervention for fine-grained image retrieval," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[60] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.

[61] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villasenor, and M. Grubinger, "The segmented and annotated iapr tc-12 benchmark," *Computer vision and image understanding*, vol. 114, no. 4, pp. 419–428, 2010.

[62] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.

SUQING ZHOU is currently an Associate Professor at The Internet of Things and Artificial Intelligence College, Fujian Polytechnic of Information Technology, Fuzhou, China. She has presided over multiple teaching and scientific research projects. She has participated in the development of the professional teaching standards of "computer network technology" specialty in China's higher vocational schools and the construction standards of practical training teaching conditions for computer specialty in China's higher vocational schools. Her research interests involve future internet architecture, network virtualization, cloud-network integration, and artificial intelligence for networking.

YU HAN graduated with a bachelor degree from Beijing University of Posts and Telecommunications in 1995, majoring in Wireless Communication. And then he received a M.Sc. degree from Beijing University of Posts and Telecommunications, majoring in Electronics and Information Engineering in 2001. Currently, he works in China Mobile Group Shandong Co., Ltd. and his main research fields include Cutting-edge Technology Research, Computer Vision, Intelligent Voice Technology, Natural Language Processing, etc.

NING CHEN (Graduate Student Member, IEEE) is currently pursuing the master's degree with the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China. His research interests include network virtualization, satellite communications, and beam hopping. Mr. Chen serves as a reviewer for several journals/conferences, such as IEEE Network, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Vehicular Technology, IEEE Access, etc.

SIYU HUANG received a master's degree from the University of Birmingham in the United Kingdom. After graduation, he has been engaged in the architecture design and development of information systems. He is currently a senior engineer in the Cognitive Intelligence Laboratory of Xiong'an Innovation Research Institute, mainly engaged in cognitive intelligence-related algorithm research and system integration development. During his tenure, he mainly participated in the research work of two horizontal subjects. In the series of subjects of flexible intelligent manufacturing technology research and development, he studied the refined 3D reconstruction of point clouds of metal structural parts, the completion of point clouds based on multi-frame registration, and the integration of point clouds. Weld seam extraction, fitting and optimization of welding trajectory. In a series of subjects based on the digital underlying control technology of panoramic dynamic images of complex urban intersections, he participated in the related research on intersection scene reconstruction based on visual information and high-precision map knowledge representation.

KOSTROMITIN KONSTANTIN IGOREVICH graduated degree (Hons.) from the Municipal Secondary School, Chelaybinsk, Russia, and entered the Physical Department of Education, Chelaybinsk State University, Chelyabinsk, with specialization in physics. He received the bachelor's and masters degrees in physics with specialization in chair physics of condensed matter from Chelyabinsk State University in 2008 and 2010, respectively, and the Ph.D. degree in physics from Chelyabinsk State University in 2013, with thesis entitled Researching of Magnetocaloric Effect in Antiferromagnetics and Twins Moving in Heusler alloys with specialization in physics of condensed matter in the Dissertation Council.

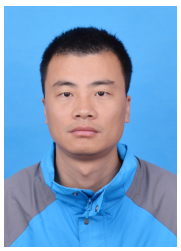
JIA LUO received the bachelor's degree from Shanghai University, China, in 2011 and the master's degree from Waseda University, Japan, in 2015. Afterwards, she received her Ph. D, degree from Université de Toulouse, France, in 2019. Currently, she is a lecturer at College of Economics and Management, Beijing University of Technology, China. Meanwhile, she is an International Research Fellow of Japan Society for the Promotion of Science. Her research interests include shop scheduling, parallel evolutionary algorithms and GPU computing.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3339581





PEIYING ZHANG (Member, IEEE) received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019. He is currently an Associate Professor with the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China. His research interests include semantic computing, future internet architecture, network virtualization, and artificial intelligence for networking.

Dr. Zhang has published multiple IEEE/ACM transactions/journal/magazine papers, such as IEEE Transactions on Industrial Informatics, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Vehicular Technology, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Network and Service Management, IEEE Network, IEEE Internet of Things Journal, and IEEE Communications Magazine. He served as the Technical Program Committee for the IEEE International Conference on Communications (ICC)'23, IEEE ICC'22, International Conference on Digital Image Processing and Pattern Recognition (DPPR) in 2021, International Conference on Information Science and Technology (ICIST) in 2022, Globecom 2022, Globecom 2021, IEEE International Conference on Communication, Networks and Satellite (COMNETSAT) in 2020, International Conference on Informatics and Computational Sciences (ICICoS) in 2022, International Workshop on Network Softwarization Techniques for IoT Applications (SoftIoT) in 2021, International Wireless Communications and Mobile Computing Conference (IWCMC) Satellite in 2020, and IWCMC-Satellite 2022.

He is the Leading Guest Editor of *Drones, Mathematics, Electronics, Wireless Communications and Mobile Computing, Mathematical Problems in Engineering*, and *Frontiers in Psychiatry*, and the editorial board of *Drones, Mobile Information Systems, Artificial Intelligence and Applications* (AIA).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4