
Unsupervised Cross-Modal Hashing Algorithms for Web Multimedia Retrieval

Yang-hao Li*, Zhao-jie Dong, Shi-song Wu,
Xuan-ang Li and Lian-yu Sha

*China Southern Power Grid Artificial Intelligence Technology Co. Ltd., Guangzhou
510700, Guangdong, China*

E-mail: yanghaoli01@163.com

**Corresponding Author*

Received 12 March 2026; Accepted 08 May 2026

Abstract

With the Web witnessing a rapid surge in multimodal data, it's becoming increasingly vital to develop efficient and budget-friendly cross-modal (CM) retrieval techniques to elevate the user experience in Web applications. Traditional hashing methods, however, often neglect the valuable semantic information hidden within the text descriptions that come with Web images. Moreover, they tend to lean heavily on supervised learning, which poses a challenge when it comes to adapting to real-world Web scenarios where annotations are often in short supply. To tackle this, this research introduces an unsupervised CM hashing algorithm for Web multimedia retrieval. By mining the semantic structure of text associated with Web images and utilizing a deep network to achieve semantic transfer from text to vision, a unified and efficient hashing learning framework is constructed. Experiments indicate that the introduced approach achieves mAP values of 0.3370 and 0.6990 with 16-bit hash codes (HCs). When the HC length is increased to 128 bits, the mAP increases to 0.3632 and 0.7575, representing an absolute improvement of 13.0% and 4.04% compared to the best performing baseline method. Further analysis shows that the semantic transfer mechanism significantly

Journal of Web Engineering, Vol. 25_6, 1085–1110.

doi: 10.13052/jwe1540-9589.2563

© 2026 River Publishers

improves the semantic representation ability of the HCs. Even in a semi-supervised setting using only 20% of labeled data, the retrieval mAP can still reach 0.8871. The method requires only a portion of image-text pairs during the training phase and supports pure image queries during the retrieval phase, achieving millisecond-level response times under Hamming distance calculation. This provides an efficient and practical solution for Web-scale multimedia retrieval.

Keywords: Unsupervised learning, cross-modal hashing, web multimedia retrieval, semantic transfer, deep neural networks.

1 Introduction

With the popularization of Web 2.0 and mobile terminals, multimodal data such as images, text, and videos have exploded [1]. How to accurately and effectively retrieve effective information from these massive heterogeneous data has emerged as a pressing issue that needs immediate resolution [2]. Currently, users are no longer content with retrieving information from just one modality, but hope to obtain semantically relevant results across modalities through a piece of text, an image, a piece of audio or video clip [3, 4]. However, the current retrieval methods have the problems of low efficiency and high cost [5]. Therefore, finding a way to realize cross-modal (CM) unsupervised retrieval is of great significance for improving retrieval efficiency and user experience.

Traditional CM hashing methods rely on hand-crafted features and shallow semantic modeling, which struggle to capture high-level semantic correlations across modalities. Recently, deep CM hashing methods have achieved significant progress by leveraging deep neural networks for feature extraction. However, these methods predominantly rely on supervised learning with large-scale annotated data, rendering them impractical in Web environments where annotations are scarce or expensive to obtain. Moreover, most existing approaches treat text and visual features independently or simply concatenate them, failing to exploit the rich semantic structure embedded in text descriptions to guide visual representation learning.

The hash method maps the key value of the data to a hash value of fixed length through a hash function, thereby determining the storage location of the data in the storage structure. In the CM scenario, this function is upgraded to a learnable semantic mapping, which compresses heterogeneous high-dimensional features into low-dimensional binary codes. It has

the advantages of both semantic preservation and fast comparison, and has been widely used in database indexing and caching systems [6]. Different from these existing approaches, this study introduces a semantic transfer mechanism that explicitly leverages text semantics as supervisory signals to guide visual feature learning, achieving unsupervised semantic transfer from text to vision through a differentiable hashing layer. Therefore, this study constructs an unsupervised CM hashing algorithm for Web multimedia retrieval based on Alex Krizhevsky Network (AlexNet) and hashing methods, aiming to improve the efficiency of Web multimedia detection. The innovation of this research lies in coupling the deep visual features of AlexNet with the semantic structure of text under completely unsupervised conditions, and establishing a directional knowledge flow from the semantic-rich text modality to the visual modality, in order to provide a theoretical foundation for related research fields in Web multimedia retrieval.

The contributions of this study are: (1) proposing an end-to-end unsupervised CM hashing framework with a semantic transfer mechanism that leverages text semantics to guide visual feature learning, which can generate compact binary codes for images and text simultaneously without any manual annotation, significantly reducing storage and computational overhead and (2) designing a joint optimization objective that enables this semantic transfer from text to vision, so that the hash code (HC) still has strong discriminative power in Web environments where annotations are extremely scarce or even missing, providing an efficient, low-cost, and practical solution for large-scale multimedia retrieval.

2 Related Work

With the rapid expansion of Internet data scale, Web multimedia retrieval has garnered significant attention from researchers. Existing methods can be categorized into graph-based methods, deep learning-based methods, and Transformer-based methods. However, these methods still face three core challenges in Web scenarios: scarce annotation, asynchronous modality, and real-time requirements.

Graph-based methods construct similarity graphs to capture semantic relationships. Zhu et al. analyzed multimodal hashing technology through literature review but did not propose substantial algorithms [7]. Liang et al. proposed self-supervised multi-granular multimodal knowledge graph contrastive hashing to address fine-grained knowledge association ignorance [8]. This method captures high-order relationships through knowledge graphs,

but its heavy reliance on graph structure quality and high computational complexity limits Web applicability. Sun et al. introduced Semantic Alignment Transformer with adversarial learning for fine-grained alignment, yet the unstable adversarial training process remains problematic [9]. The fundamental limitation of graph-based methods is their dependence on pre-defined similarity graphs, which becomes unreliable with scarce Web annotations.

Deep learning-based methods primarily employ convolutional neural network architectures with different loss designs. Han et al. proposed supervised hierarchical online hashing using multi-level semantic structures, but its supervised nature requires abundant labels unavailable in Web scenarios [10]. Li et al. introduced lightweight weakly supervised hashing with pseudo-label generation [11]. While reducing annotation dependency, pseudo-label quality directly affects performance with unresolved error propagation. Liu et al. proposed scalable unsupervised hashing for CM consistency, but ignored modality-specific characteristics and suffered from slow convergence [12]. Song et al. developed sorted distribution-preserving hashing to utilize complete ranking information, yet remained sensitive to out-of-distribution semantics [13]. These methods essentially differ in loss functions (from pairwise to distribution-preserving) but commonly suffer from insufficient semantic capture and weak CM interaction.

Transformer-based methods leverage self-attention for global dependency modeling. Wang et al. proposed a multi-granularity Transformer framework, but standard Transformer's quadratic complexity is unsuitable for real-time retrieval [14]. Li et al. introduced similarity attention fusion hash for adaptive matrix fusion, yet lacked theoretical justification for fusion strategies [15]. Sun et al. developed dual self-propelled hash with self-supervised learning for robustness, but the dual-network structure doubled computational overhead [16]. While capturing long-range relationships better than convolutional neural networks, these methods fail to address Web modality asynchrony and introduce unacceptable latency.

Recent advances explore federated and generative approaches. To address privacy and communication costs in centralized training, Zuo et al. proposed privacy-enhanced prototype-based federated CM hashing with encrypted prototype transmission. However, federated aggregation introduces communication overhead unsuitable for real-time Web retrieval [17]. To bridge the discrete-continuous gap, Tu et al. proposed unified generative hashing with Uniform Mixture-of-Expert Decoder for precise alignment, but iterative optimization affects retrieval speed [18].

The above research indicates that existing methods systematically fail to address Web scenario challenges: scarce annotation with error propagation risks, asynchronous modality with misalignment issues, and real-time requirements conflicting with complex architectures. Therefore, this study proposes an unsupervised CM hashing algorithm based on AlexNet, mining textual semantic structure for zero-label semantic transfer and constructing efficient differentiable hashing to simultaneously improve retrieval speed and storage efficiency.

3 Research Methods

3.1 Unsupervised CM Hashing Framework based on Semantic Transfer

3.1.1 Hash algorithm framework and its unsupervised deep hash module

In the context of the information explosion era, efficient Web retrieval is of paramount importance. To improve the speed of Web retrieval, this study designs an unsupervised CM hashing framework based on AlexNet semantic transfer, as shown in Figure 1.

As shown in Figure 1, the framework first extracts visual features from the original image through a multi-layer convolutional network. Then, the unsupervised semantic information from the text side is injected into the visual channel with the help of the semantic transfer module, so that the image features retain their own similarity and text semantics at the same time. Then, the network jointly generates binary HC and reusable hash function at the same end to complete the training. Finally, the obtained function can generate compact encoding for any query image in real time, realizing fast retrieval

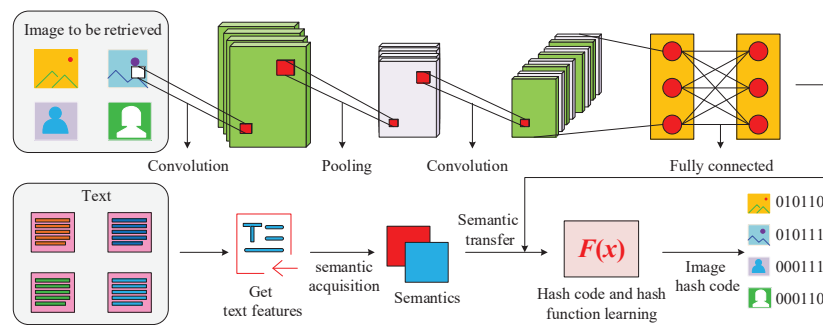


Figure 1 Unsupervised CM hashing framework based on AlexNet semantic transfer.

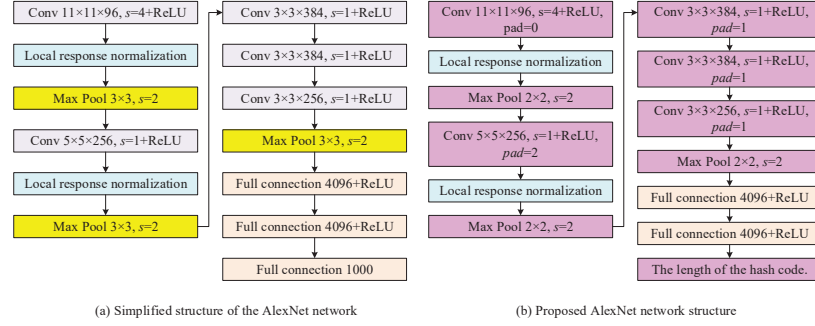


Figure 2 Simplified structure of the basic AlexNet and the proposed AlexNet structure.

of Web images. The multi-layer convolutional network in this framework is based on AlexNet. AlexNet is a deep convolutional neural network architecture with advantages such as simple structure and easy transfer, and has been widely used in image recognition and medical diagnosis [19]. The simplified structure of the basic AlexNet and the proposed AlexNet structure are shown in Figure 2 [20].

In Figure 2, Conv represents a convolutional layer, s represents the convolution stride, pad is the number of zero-pixel layers added around the input feature map, and ReLU represents the ReLU activation function. In the convolutional layer, “ $11 \times 11 \times 96$ ” represents “size $\times$ size $\times$ channel,” and max pooling “ 3×3 ” means taking the maximum value within a 3×3 local window as the output. As can be seen from Figures 2(a) and 2(b), the proposed AlexNet network structure changes the pooling kernel size from 3×3 to 2×2 , reducing the feature map downsampling rate and preserving more spatial information. The convolutional layer padding strategy is explicitly defined: $pad = 0$ for Conv1, $pad = 2$ for Conv2, and $pad = 1$ for Conv3-5, precisely controlling the output size of each layer. The output dimension of the fully connected layer is changed from 1000 classes to HC length, adapting to hash retrieval tasks. Through multiple experiments setting the above parameters, the feature map size can be reasonably reduced while ensuring feature representation capabilities, thus achieving unsupervised deep hashing based on AlexNet. The proposed AlexNet acts as a visual feature extractor in the semantic transformation mechanism. By adjusting the size of the pooling kernel and the padding strategy, this network can better preserve spatial information. The first seven layers extract high-level visual features from the image, while the eighth layer aligns these features with the text semantics, thereby achieving semantic transformation from text to vision.

The proposed AlexNet adopts a hierarchical training strategy. The first seven layers are initialized using the pre-trained weights from ImageNet and fine-tuned with a learning rate of 0.0001 to prevent overfitting, while the eighth layer is randomly initialized with a learning rate of 0.001. The network is trained using batch gradient descent with a batch size of 64 for 400 iterations. In an unsupervised condition, the semantic similarity matrix constructed from the text topic distribution serves as a soft label for supervising visual hash learning without the need for manual annotation. In a semi-supervised condition (a certain number of data is labeled), the labeled part is strictly supervised with real labels, while the remaining 80% relies on the semantic similarity matrix for weak supervision.

3.1.2 K-means-based unsupervised semantic learning method

In the area of unsupervised semantic learning of text, this study proposes an unsupervised semantic learning method based on K-means clustering (K-means). The implementation process of this approach is presented in Figure 3.

Figure 3 illustrates the process of this unsupervised semantic learning method. First, it receives text semantic feature $X^{(2)} = \{X_i^{(2)}\}_{i=1}^N$ used for training as input, where the sample size is N . Simultaneously, it sets the number of cluster columns K , parameter σ , and sparse constraint weight λ . Second, it calculates the similarity matrix $A = \{a_{ij} | 1 \leq i \leq N, 1 \leq j \leq N\}$, where a_{ij} is the element in the row i and column j of the matrix A , and the similarity calculation of a_{ij} can be expressed by Equation (1):

$$a_{ij} = e^{\frac{-\cos^2(x_i^{(2)}, x_j^{(2)})}{2\sigma^2}} \quad (1)$$

where $x_i^{(2)}$ and $x_j^{(2)}$ are the i th and j th text features, and $\cos(\cdot, \cdot)$ is the cosine similarity function. Next, to avoid the influence of the similarity of

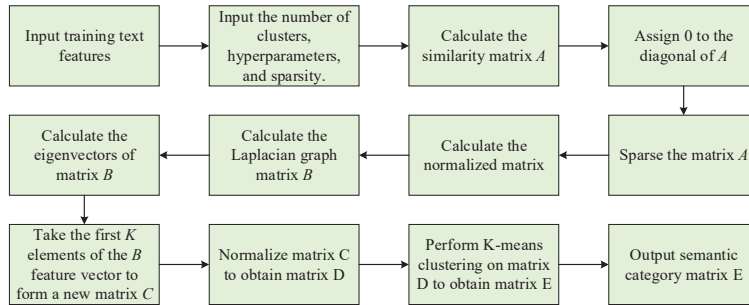


Figure 3 Unsupervised semantic learning method based on K-means.

the elements themselves, the elements on the diagonal of the matrix A are assigned values, i.e. $a_{ij} = 0$. Meanwhile, to improve the clustering effect, the matrix A is sparsified, which can be expressed by Equation (2):

$$A(a_{ij} < \lambda \bar{a}) = 0 \quad (2)$$

where $\bar{a}$ is the average value of all elements in the matrix A . Next, the diagonal elements in the matrix A are normalized to obtain the degree matrix Q , which can be represented by Equation (3):

$$Q(i, j) = \sum_{j=1}^N a_{ij} \quad (3)$$

After obtaining the degree matrix Q , its Laplace graph matrix B can be calculated, which can be expressed by Equation (4):

$$B = Q^{\frac{1}{2}} A Q^{\frac{1}{2}} \quad (4)$$

Then, the eigenvalues of the matrix B and their corresponding eigenvectors are calculated, arranged in descending order, and the top-ranked K eigenvectors $c_1, c_2, \dots, c_K$ are selected to form a matrix $C = (c_1, c_2, \dots, c_K)$. Finally, the elements in each row of the matrix C are normalized to obtain the normalized eigenvalue matrix E , which can be represented by Equation (5):

$$E_{ij} = \frac{C_{ij}}{(\sum_{j=1}^N C_{ij}^2)^{\frac{1}{2}}} \quad (5)$$

Finally, each row in the matrix E is taken as a point in the K -dimensional space, and the K-means algorithm is used for clustering to obtain the category of each text feature. Thus, the category label vector $R = \{r_i\}_{i=1}^N$ of the text feature is obtained. Among them, K-means is an unsupervised learning clustering algorithm with advantages such as simplicity, intuitiveness and ease of implementation, and has been widely used in large-scale data processing and other fields. With the number of clusters as 3, the basic idea and calculation process of the K-means algorithm are presented in Figure 4 [21].

Figure 4(a) illustrates that the K-means algorithm aims to partition the dataset into three clusters, ensuring each data point is assigned to the nearest cluster center. Through iterative adjustments of the cluster centers' positions, K-means continually enhances cluster density, resulting in clusters that are as tightly grouped and distinct from one another as possible. Figure 4(b)

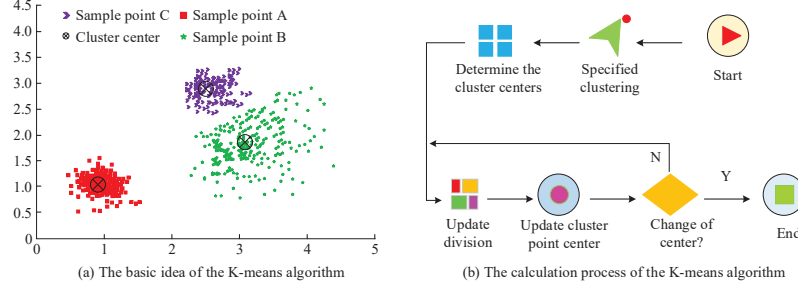


Figure 4 Basic idea and calculation of the K-means.

shows the K-means algorithm's calculation process. Initially, the algorithm randomly chooses K sample points to serve as the starting cluster centers. Next, it computes the distance from each sample to every center and allocates it to the closest cluster. Next, it recalculates the centroid of each cluster (i.e., the mean of all samples in that cluster) as the new cluster center. Then, if the cluster centers no longer change or the change is less than a threshold, the algorithm terminates; otherwise, it returns to step 1 to continue iterating. Finally, output the cluster centers and the class label of each sample.

3.2 CM Semantic Alignment and Hashing Optimization

Based on the K-means unsupervised semantic learning method and the proposed AlexNet, image and semantic features are extracted. A CM semantic alignment and hash optimization method is then proposed. This method first calculates the similarity matrix $H = \{h_{ij}\}$ for each data point using the semantic learning result R , and $h_{ij} \in \{0, 1\}$. If data points Z_i and Z_j are similar, then $h_{ij} = 1$; otherwise $h_{ij} = 0$. It can be expressed by Equation (6):

$$h_{ij} = \begin{cases} 1, & r_i = r_j \\ 0, & \text{Others} \end{cases} \quad (6)$$

Secondly, if the binary code of all images is given as $Y = \{y_i\}_{i=1}^N \in \{-1, 1\}^{L \times N}$, where L is the hash length, then the optimal HC matrix is found by the Maximum Likelihood Estimation (MLE) algorithm, which maximizes the probability of the observed similarity matrix H appearing, and can be expressed by Equation (7):

$$p(h_{ij}|Y) = \begin{cases} \sigma(o_{ij}), & h_{ij} = 1 \\ 1 - o_{ij}, & h_{ij} = 0 \end{cases} \quad (7)$$

where o_{ij} is the inner product of HCs, i.e., Hamming similarity. Therefore, by transforming MLE into a specific optimization objective function, the Hamming distance between two independent data points that satisfy similarity is minimized and, vice versa, the image hashing learning ability of the algorithm is improved. This can be expressed by Equation (8):

$$\min_Y l_1 = -\log p(H|Y) = - \sum_{h_{ij} \in H} (h_{ij} o_{ij} - \log(1 + e^{o_{ij}})) \quad (8)$$

where l_1 is the objective function. However, Equation (8) is a discrete optimization problem, and it is an NP-hard problem with high computational complexity, making it difficult to solve directly. Therefore, this study optimized it. Then, auxiliary variable u_i is introduced to relax it, and the objective function can be expressed by Equation (9):

$$\begin{cases} \min_{Y,U} l_2 = - \sum_{h_{ij} \in H} (h_{ij} T_{ij} - l \log(1 + e^{T_{ij}})) \\ s.t. \ u_i = y_i, \quad \forall i = 1, 2, \dots, N \\ \quad u_i \in \{-1, 1\}^{L \times 1}, \quad \forall i = 1, 2, \dots, N \\ \quad y_i \in \{-1, 1\}^L, \quad \forall i = 1, 2, \dots, N \end{cases} \quad (9)$$

where l_2 is the relaxed objective function, U is the matrix of continuous relaxed variables, and T_{ij} is the inner product similarity of the continuous variables, where $T_{ij} = \frac{1}{2} u_i^T u_j$ and $U = \{u_i\}_{i=1}^N$. Then, the discretized HC can be obtained through the formula $y_i = \text{sgn}(u_i)$. To improve the solution speed of Equation (9), a regularization term coefficient γ is introduced to transform its constraint condition into an optimization regularization problem, which can be expressed by Equation (10):

$$\min_{Y,U} l_2 = - \sum_{h_{ij} \in H} (h_{ij} T_{ij} - l \log(1 + e^{T_{ij}})) + \gamma \sum_{i=1}^N \|y_i - u_i\|_2^2 \quad (10)$$

Finally, to ensure that the hash representation generated by the AlexNet corresponds to the variables in Equation (10), thereby embedding the AlexNet into the entire hash learning framework, let $u_i = W^T \psi(X_i^{(1)}; \theta) + \chi$, where W is the weight matrix, χ is the offset vector, θ is the parameter of the AlexNet, $X_i^{(1)}$ is the image input to the AlexNet, and $\psi(X_i^{(1)}; \theta)$ is the output of the AlexNet. Then, by connecting the acquired image features

and semantic transfer results using χ and θ , and substituting u_i into Equation (10), CM semantic alignment can be achieved, which can be expressed by Equation (11):

$$\begin{aligned} \min_{Y, U, \chi, \theta} l_2 = & - \sum_{h_{ij} \in H} (h_{ij} T_{ij} - l \log(1 + e^{T_{ij}})) \\ & + \gamma \sum_{i=1}^N \|y_i - (W^T \psi(X_i^{(1)}; \theta) + \chi)\|_2^2 \end{aligned} \quad (11)$$

Thus, end-to-end fusion of semantic transfer and hashing processes (including function construction and code value generation) is achieved. Finally, hash optimization is achieved by solving the problem based on an alternating optimization strategy. This method obtains the HC by fixing parameters θ , χ , and W and discretizing the symbol function y_i . It can be expressed by Equation (12):

$$y_i = \text{sgn}(W^T \psi(X_i^{(1)}; \theta) + \chi) \quad (12)$$

The learning of θ , χ , and W is achieved by backpropagation gradient calculation. First, the gradient of the loss function of u_i is solved, which can be expressed by Equation (13):

$$\frac{\partial \tau}{\partial u_i} = \frac{1}{2} \sum_{j: h_{ij}=H} (\beta_{ij} - h_{ij}) u_j + \frac{1}{2} \sum_{j: h_{ij}=H} (\beta_{ij} - h_{ij}) u_j + 2\gamma(u_i - y_i) \quad (13)$$

where β_{ij} is the prediction probability, and $\beta_{ij} = \sigma(\frac{1}{2} u_i^T u_j)$. Secondly, taking Equation (13) as the output error of hash learning, the gradients of θ , χ , and W can be calculated, which can be expressed by Equation (14):

$$\begin{cases} \frac{\partial \tau}{\partial W} = \frac{\partial \tau}{\partial u_i} \cdot \frac{\partial u_i}{\partial W} = \psi(X_i^{(1)}; \theta) \left(\frac{\partial \tau}{\partial u_i} \right)^T \\ \frac{\partial \tau}{\partial \chi} = \frac{\partial \tau}{\partial u_i} \cdot \frac{\partial u_i}{\partial \chi} = \frac{\partial \tau}{\partial u_i} \\ \frac{\partial \tau}{\psi(X_i^{(1)}; \theta)} = \frac{\partial \tau}{\partial u_i} \cdot \frac{\partial u_i}{\psi(X_i^{(1)}; \theta)} = W \frac{\partial \tau}{\partial u_i} \end{cases} \quad (14)$$

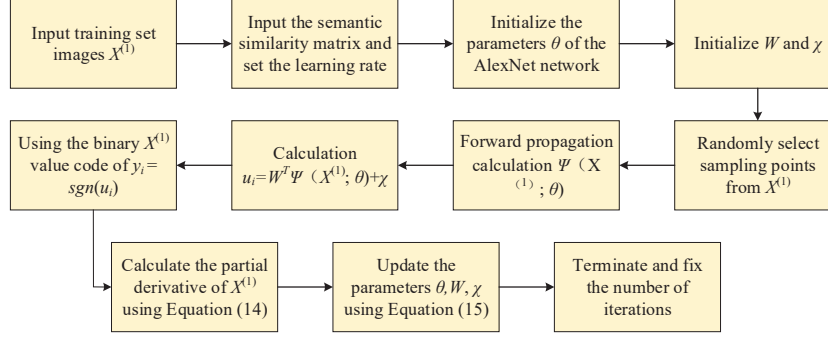


Figure 5 Modal semantic alignment and hash learning optimization methods.

Then, the updated θ , χ , and W can be obtained according to Equations (13) and (14), which can be expressed by Equation (15):

$$\begin{cases} W^+ = W - \vartheta \frac{\partial u_i}{\partial W} = W - \vartheta \psi(X_i^{(1)}; \theta) \left(\frac{\partial \tau}{\partial u_i} \right)^T \\ \chi^+ = \chi - \vartheta \frac{\partial \tau}{\partial \chi} = \chi - \vartheta \frac{\partial \tau}{\partial u_i} \\ \psi(X_i^{(1)}; \theta)^+ = \psi(X_i^{(1)}; \theta) - \chi W \frac{\partial \tau}{\partial u_i} \end{cases} \quad (15)$$

where W^+ , χ^+ , and $\psi(X_i^{(1)}; \theta)^+$ are the updated θ , χ , and W , and ϑ is the learning rate. Thus, the modal semantic alignment and hash learning optimization method is completed, as shown in Figure 5.

As shown in Figure 5, this method first inputs the training image $X^{(1)}$ and similarity matrix H , sets the learning rate ϑ , and initializes the parameters θ , χ , and W . Then, it randomly selects sample points from $X^{(1)}$ and calculates u_i . Next, the binary code of $X^{(1)}$ is calculated using the formula $y_i = \text{sgn}(u_i)$. Then, forward propagation is performed using Equation (14) to calculate the partial derivative of the loss function with respect to, u_i i.e., the gradient of the loss function. Then, backward propagation is performed using Equation (15) to update the parameters θ , χ , and W . Iteration is performed until the preset iteration count is reached to complete the training. Finally, after training is completed, in order to realize the algorithm's retrieval of unknown images, a hash function y_q is designed and the hash algorithm is optimized, which can be expressed by Equation (16):

$$y_q = s(X_q) = \text{sgn}(W^T \psi(X_q; \theta) + \chi) \quad (16)$$

where X_q is any image, and $X_q \notin X^{(1)}$. Therefore, when retrieving images on the Web, Equation (16) can be used to hash the image and then quickly compare it with samples in the database in the binary space to return the nearest neighbor matching result.

4 Experimental Results

4.1 Experimental Setup

To confirm the capability of the introduced unsupervised CM hashing algorithm, a performance comparison analysis was conducted with the UKD-SS, DJSRH, and CMFH algorithms. The data sources were the publicly available datasets: the WiKi and the NUS-WIDE. The WiKi dataset can be downloaded from: <http://www.svcl.ucsd.edu/projects/crossmodal/> and the NUS-WIDE dataset from: <https://opendatalab.com/OpenDataLab/NUS-WIDE/tree/main>. The WiKi dataset contains 2866 image-text pairs, and the NUS-WIDE dataset used 5200 images after filtering. During data preprocessing, duplicate images were removed. The text portion used the Latent Dirichlet Allocation algorithm to extract topic features, and the image portion used the original pixels as input. The datasets were divided into training and test sets in an 8:2 ratio. After multiple experiments and optimizations, the algorithm was optimized on the WiKi dataset with 10 spectral clusters, a Gaussian kernel parameter of 0.1 for the similarity matrix, a sparsity regularization term of 1, and HC length ranging from 16 to 128 bits. The search size increased in increments of 100 from 100 to 1000. On the NUS-WIDE dataset, the algorithm also used 10 spectral clusters, a hyperparameter of 1, and a sparsity of 0.8. The input image size was 227×227 , and the original image was used directly for retrieval. Text descriptions were introduced during the training phase, and visual feature extraction was optimized through a CM alignment mechanism. The algorithm framework iterated 400 times. The network architecture used pre-trained weights from AlexNet on the ImageNet dataset for the first seven layers, and the learning rate for these seven layers was set to 0.0001 to suppress overfitting. Experimental comparison metrics included mean average precision (mAP) and average precision (AP). The environment settings are presented in Table 1.

To verify that the proposed method has optimal parameter settings, a sensitivity analysis was conducted on the WiKi dataset, using retrieval speed as the indicator. The analysis outcomes are presented in Table 2.

Table 2 shows that when the values of σ , K , and λ were 0.1, 10, and 1, respectively, and the retrieval time was 7.1 ms, which was better than other

Table 1 Experimental environment configuration

Parameter Names	Parameter
CPU	Intel Core i9-14900K
Main frequency	6.0GHz
Internal memory	32GB
Hard disk capacity	500TB
Operating system	Windows 10 64
GPU	NVIDIA GeForce RTX 4090
MATLAB version	MATLAB 2021b

Table 2 Sensitivity analysis results in the WiKi dataset

Parameter Combination	$K=2$	$K=4$	$K=6$	$K=8$	$K=10$	$K=12$
$\sigma = 0.01, \lambda = 0$	9.2 ms	8.8 ms	8.5 ms	8.3 ms	8.1 ms	8.6 ms
$\sigma = 01, \lambda = 1$	8.9 ms	8.5 ms	8.2 ms	8.0 ms	7.8 ms	8.3 ms
$\sigma = 01, \lambda = 2$	8.7 ms	8.3 ms	8.0 ms	7.8 ms	7.6 ms	8.1 ms
$\sigma = 1, \lambda = 0$	8.5 ms	8.1 ms	7.8 ms	7.6 ms	7.4 ms	8.5 ms
$\sigma = 1, \lambda = 1$	8.2 ms	7.8 ms	7.5 ms	7.3 ms	7.1 ms	8.2 ms
$\sigma = 1, \lambda = 2$	7.9 ms	7.5 ms	7.8 ms	7.6 ms	7.9 ms	8.4 ms
$\sigma = 1, \lambda = 0$	8.8 ms	8.4 ms	8.1 ms	7.9 ms	7.7 ms	7.6 ms
$\sigma = 1, \lambda = 1$	8.4 ms	8.0 ms	7.7 ms	7.5 ms	7.3 ms	7.2 ms
$\sigma = 1, \lambda = 2$	8.1 ms	7.7 ms	7.4 ms	7.5 ms	7.8 ms	8.7 ms
$\sigma = 10, \lambda = 0$	9.5 ms	9.1 ms	8.8 ms	8.5 ms	8.7 ms	8.6 ms
$\sigma = 10, \lambda = 1$	9.3 ms	8.9 ms	8.6 ms	8.3 ms	8.1 ms	8.0 ms
$\sigma = 10, \lambda = 2$	9.0 ms	8.6 ms	8.3 ms	8.0 ms	7.8 ms	7.7 ms
$\sigma = 100, \lambda = 0$	1.5 ms	1.1 ms	9.8 ms	9.5 ms	9.3 ms	9.1 ms
$\sigma = 100, \lambda = 1$	1.2 ms	9.8 ms	9.5 ms	9.2 ms	9.0 ms	8.8 ms
$\sigma = 100, \lambda = 2$	9.9 ms	9.5 ms	9.2 ms	8.9 ms	8.7 ms	8.5 ms

parameter settings. These results indicated that the proposed method achieved optimal performance, and the parameter settings were effective.

4.2 Overall Performance Comparison and Analysis

After setting the parameters and experimental environment, the study first conducted a comparative experiment on two datasets using mAP as the metric for each algorithm in this environment. The experimental outcomes are presented in Figure 7.

As shown in Figure 6(a), with a short HC of 16 bits, the introduced unsupervised CM hashing algorithm achieved mAP values of 0.3370 and 0.6990 in the WiKi and NUS-WIDE datasets. When the HC length increased

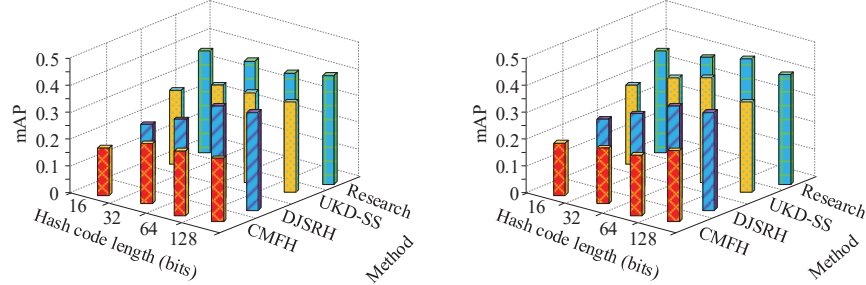


Figure 6 Various methods in different mAP comparison results of the dataset.

to 128 bits, the mAP values increased to 0.3632 and 0.7575, respectively. Compared to the comparative methods, these represented absolute improvements of 13.0% and 4.04%. Figure 6(b) shows that in the NUS-WIDE dataset, the proposed hashing algorithm achieved higher mAP values than the comparative algorithms for all hash lengths. A higher mAP indicates that the retrieval results are closer to the actual results. These results demonstrated that, from the perspective of mAP, the proposed method outperformed the comparative methods. The precision comparison results of each algorithm in the WiKi dataset are shown in Figure 8.

As shown in Figure 7, in the WiKi dataset, the precision of all algorithms reached its lowest point when the retrieval volume was 1000. Figures 7(a)–7(d) show that when the HC length was 16 bits, 32 bits, 64 bits, and 128 bits, the precision of the proposed hash algorithm was 0.098, 0.130, 0.120, and 0.140, respectively, all higher than that of the UKD-SS algorithm (0.087, 0.089, 0.098, 0.099), the DJSRH algorithm (0.047, 0.043, 0.076, 0.064), and the CMFH algorithm (0.019, 0.004, 0.012, 0.014). These results indicated that the proposed hash algorithm outperformed the compared algorithms in terms of retrieval accuracy in the WiKi dataset. Figure 9 shows the comparison of precision of each algorithm on the NUS-WIDE dataset.

As shown in Figure 8, in the NUS-WIDE dataset, the precision of all algorithms steadily decreased with increasing retrieval volume. Figures 8(a)–8(d) reveal that when the HC length was 16 bits, 32 bits, 64 bits, and 128 bits, the precision curves of the proposed hash algorithm were higher than those of the UKD-SS, DJSRH, and CMFH algorithms. These results demonstrated that the proposed hash algorithm outperformed the compared algorithms in the NUS-WIDE dataset and exhibited good generalization ability. To further verify the performance of the proposed hash algorithm, the study conducted a performance comparison analysis experiment on the NUS-WIDE dataset

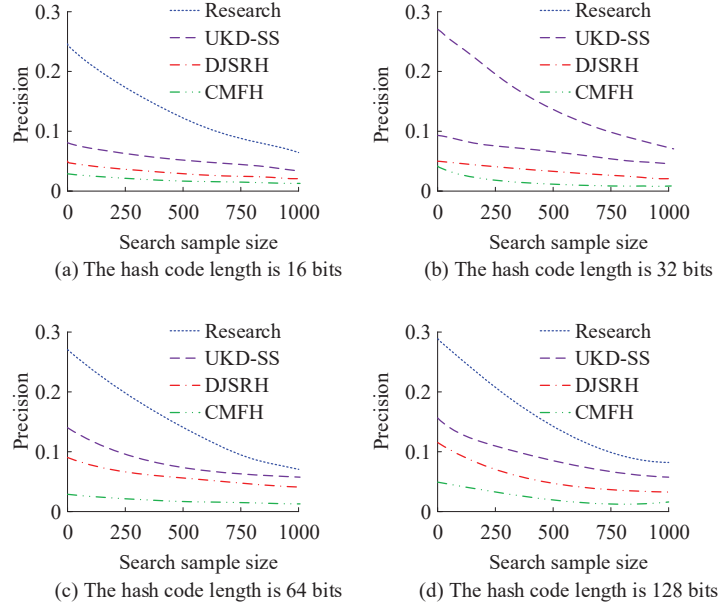


Figure 7 Comparison results of the precision rates of each algorithm on the WiKi dataset.

Table 3 Performance comparison and analysis results of each algorithm

Data Set	HC Length	Research	UCCH [22]	CIRH [23]
WiKi	16 bits	0.3370	0.3321	0.3284
	32 bits	0.3452	0.3386	0.3367
	64 bits	0.3521	0.3446	0.3421
	128 bits	0.3632	0.3515	0.3501
NUS-WIDE	16 bits	0.6990	0.6847	0.6814
	32 bits	0.7215	0.7082	0.7036
	64 bits	0.7468	0.7314	0.7288
	128 bits	0.7575	0.7453	0.7413

and the WiKi dataset, using the optimal mAP of text-image as the evaluation metric. The comparison algorithms were the UCCH [22] algorithm and the CIRH [23] algorithm. The open-source code address of the UCCH algorithm is: <https://github.com/penghu-cs/UCCH>. The open-source code address of the CIRH algorithm is: <https://github.com/XizeWu/CIRH>. The experimental results are shown in Table 3.

As shown in Table 3, for different HC lengths, in the WiKi dataset, when the hash length was 128 bits, the mAP of the proposed algorithm was 0.3632,

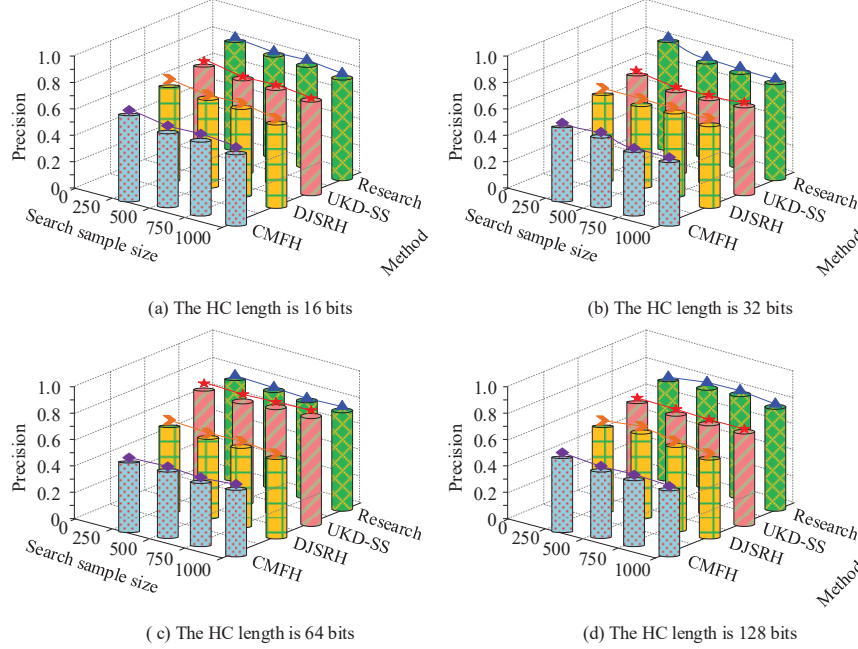


Figure 8 Comparison results of the precision rates of each algorithm on the NUS-WIDE dataset.

which was approximately 1.22% and 1.32% higher than that of UCCH and CIRH, respectively. On the NUS-WIDE dataset, the mAP of the proposed algorithm at 128 bits was 0.7575, which exceeded the UCCH and CIRH algorithms by 1.68% and 2.09%, respectively. The above results indicated that the superiority of the proposed algorithm in the unsupervised CM hashing task was evident, and it could effectively learn more discriminative CM feature representations.

4.3 Validation of the Effectiveness of the Semantic Transfer Mechanism

To verify the effectiveness of the semantic transfer mechanism, the study first conducted a query experiment with pure image queries under the 0% annotation condition. On the WiKi dataset, with an HC length of 128 bits, the mAP value was 0.948. On the NUS-WIDE dataset, with an HC length of 128 bits, the mAP value was 0.936. Both were superior to the comparison algorithms. Both were superior to UKD-SS, DJSRH, and CMFH.

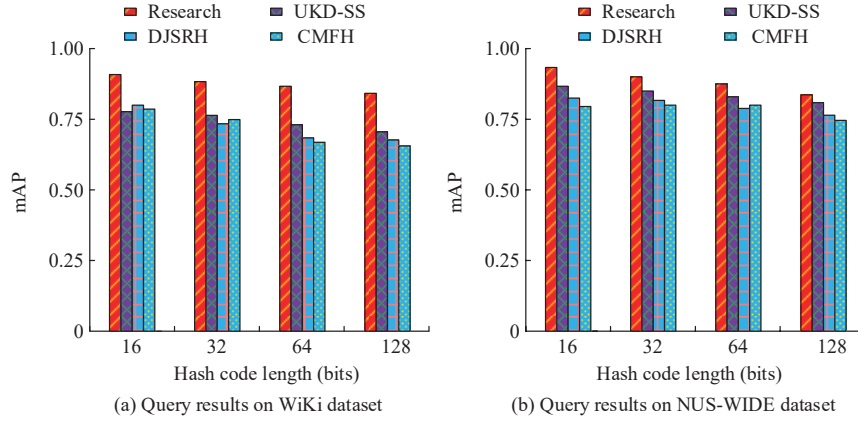


Figure 9 Retrieval mAP results under the 20% annotation (semi-supervised) condition.

Subsequently, query experiments were conducted under conditions of only 20% annotation (semi-supervised) and pure image queries. It should be noted that the proposed method was, by default, completely unsupervised (without any labeling). The 20% labeled experiment was conducted merely for ablation analysis to verify the effectiveness of the semantic transfer mechanism, rather than as a standard operational mode. The retrieval mAP results under the 20% annotation (semi-supervised) condition are shown in Figure 9.

As shown in Figure 9(a), under the 20% annotation (semi-supervised) condition in the WiKi dataset, the retrieval mAP values of all algorithms decreased with the increase of HC length. When the HC length was 128 bits, the mAP values of the proposed hash algorithm, UKD-SS algorithm, DJSRH algorithm, and CMFH algorithm were 0.8871, 0.7476, 0.7328, and 0.7251, respectively. Among them, the proposed hash algorithm had the highest mAP value. Figure 9(b) shows that under the 20% annotation (semi-supervised) condition in the NUS-WIDE dataset, the proposed hash algorithm had a higher mAP than the comparative algorithms under different HC lengths. These results indicated that the proposed hash algorithm still had good retrieval performance under the 20% annotation (semi-supervised) condition, outperforming the comparative algorithms. Each algorithm only required partial image-text pairs during the training phase and supported pure image queries during the retrieval phase. The response time comparison results under Hamming distance calculation are shown in Figure 10.

As shown in Figures 10(a) and 10(b), only a subset of image-text pairs were required during the training phase. The proposed hash algorithm

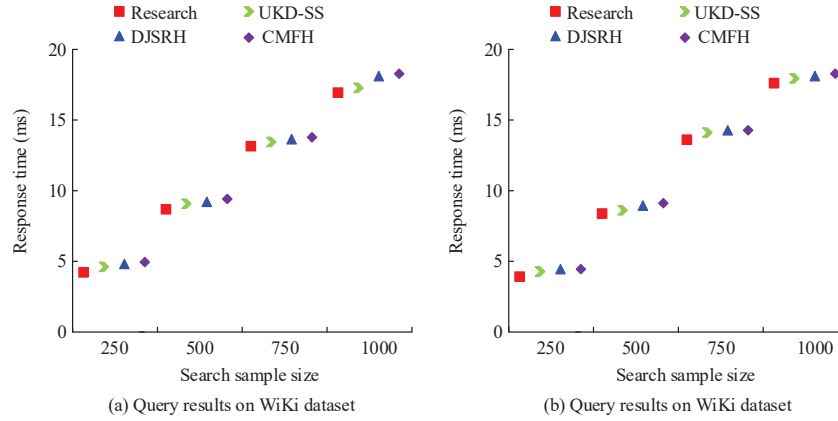


Figure 10 Comparison results of pure image retrieval response times.

supported pure image queries during the retrieval phase. The response time under Hamming distance calculation remained within the range of 4 ms to 18 ms as the retrieval volume increased, achieving millisecond-level response times. Furthermore, the retrieval results on both datasets were superior to the UKD-SS, DJSRH, and CMFH algorithms. In conclusion, the proposed semantic transfer mechanism exhibited excellent performance and outperformed the compared algorithms. The performance variations across datasets can be attributed to their inherent characteristics. The Wiki dataset featured high text-image correlation with well-structured Wikipedia articles, enabling more effective semantic guidance for visual feature learning. In contrast, NUS-WIDE relied on user-generated tags with looser semantic associations, which limited the supervisory capacity of textual semantics. Additionally, Wiki's smaller category space (10 classes) facilitated more accurate unsupervised semantic clustering compared to NUS-WIDE's larger label space (81 classes). These factors collectively explained the observed performance gap between the two datasets.

5 Discussion

This study conducted comparative experiments on the unsupervised CM hashing framework based on AlexNet semantic transfer and verified the effectiveness of the semantic transfer mechanism. In the retrieval accuracy comparison experiment, the unsupervised CM hashing framework proposed in this study achieved mAP values of 0.3632 and 0.7575 under the 128-bit HC

condition in the WiKi dataset and NUS-WIDE dataset, which is an absolute improvement of 13.0% and 4.04% compared with the UKD-SS, DJSRH, and CMFH algorithms. This result showed that the unsupervised semantic learning method based on K-means effectively captured the latent semantic structure of the text modality, and provided a reliable supervision signal for image hashing learning through the semantic category matrix optimized by spectral clustering and sparse constraints. The performance variations across datasets could be attributed to their inherent characteristics. The WiKi dataset featured high text-image correlation with well-structured Wikipedia articles, enabling more effective semantic guidance for visual feature learning. In contrast, NUS-WIDE relied on user-generated tags with looser semantic associations, which limited the supervisory capacity of textual semantics. Additionally, WiKi's smaller category space (10 classes) facilitated more accurate unsupervised semantic clustering compared to NUS-WIDE's larger label space (81 classes). These factors collectively explained the observed performance gap between the two datasets. At the same time, the semantic transfer mechanism based on AlexNet injected unsupervised semantic information from the text side into the visual channel, so that the image features could fuse text semantics while maintaining their own similarity, thereby optimizing the semantic representation ability of the HC. In the precision comparison experiment, when the number of searches in the WiKi dataset was 1000, the precision of the proposed hash algorithm under 16-bit, 32-bit, 64-bit, and 128-bit HCs were 0.098, 0.130, 0.120, and 0.140, respectively, all of which were significantly higher than the comparison algorithm.

In the semantic transfer mechanism effectiveness verification experiment, the proposed hash algorithm achieved mAP values of 0.8871 and the corresponding optimal values in the WiKi dataset and NUS-WIDE dataset under the condition of 20% labeling, which was significantly better than UKD-SS's 0.7476, DJSRH's 0.7328, and CMFH's 0.7251. This result showed that the HC matrix optimized by MLE could make full use of limited labeling information and combine pseudo-labels of unsupervised semantic learning to achieve efficient semi-supervised hash learning. By using fixed parameter discretization and backpropagation gradient calculation, the convergence and stability of the algorithm under partially labeled data are guaranteed. In addition, the response time of the proposed hash algorithm under Hamming distance calculation was stable in the range of 4 ms to 18 ms as the retrieval volume increases, which was better than UKD-SS, DJSRH and CMFH algorithms. The findings demonstrated that the proposed AlexNet-based network structure effectively reduced computational complexity while ensuring

feature representation capabilities. The binary representation of HCs and the efficient calculation of Hamming distance enable large-scale retrieval to be completed in milliseconds, meeting the real-time requirements of Web-scale multimedia retrieval. This performance is comparable to the deep hash fast retrieval scheme proposed by Wu et al., indicating that the method achieves a good balance between retrieval efficiency and accuracy [24]. The above findings demonstrated that the proposed hash algorithm was effective and real-time. Although this study meets the real-time and accuracy requirements of Web retrieval, in the future, new architectures such as Vision Transformer combined with lightweight methods can be used to further improve retrieval accuracy.

6 Conclusion

To address the challenge of CM retrieval in Web environments with scarce annotated data, an unsupervised hashing framework was developed that transfers semantic knowledge from text to visual modalities, thereby eliminating reliance on manual labeling. Extensive experiments conducted on standard benchmarks demonstrated that this approach achieved superior retrieval precision and mean average precision compared to existing baseline methods. Notably, the framework maintained effective performance even with limited labeled data and achieved response times suitable for real-time applications. These results established that semantic transfer constituted a viable mechanism for large-scale multimedia retrieval without extensive supervision.

References

- [1] Zhu Y, Wu Y, Sebe N, Yan Y. Vision+ x: A survey on multimodal learning in the light of data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 9102–9122. DOI:10.1109/TPAMI.2024.3420239.
- [2] Han Z, Azman A B, Mustaffa M R B, Khalid, F B. Cross-modal retrieval: A review of methodologies, datasets, and future perspectives. *IEEE Access*, 2024, 12(1): 115716–115741. DOI:10.1109/ACCESS.2024.3444817.
- [3] Li T, Kong L, Yang X, Wang B, Xu J. Bridging modalities: A survey of cross-modal image-text retrieval. *Chinese Journal of Information Fusion*, 2024, 1(1): 79–92. DOI:10.62762/CJIF.2024.361895.

- [4] Ma X, Yang M, Li Y, Hu P, Lv J, Peng X. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 2024, 33(1): 2587–2598. DOI:10.1109/tip.2024.3374221.
- [5] Wang Z, Xu X, Wei J, Xie N, Yang Y, Shen H T. Semantics disentangling for cross-modal retrieval. *IEEE Transactions on Image Processing*, 2024, 33(1): 2226–2237. DOI:10.1109/tip.2024.3374111.
- [6] Hu Z, Cheung Y M, Li M, Lan W. Cross-modal hashing method with properties of hamming space: A new perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 7636–7650. DOI:10.1109/tpami.2024.3392763.
- [7] Zhu L, Zheng C, Guan W, Li J, Yang Y, Shen H T. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 36(1): 239–260. DOI:10.1109/tkde.2023.3282921.
- [8] Liang M, Du J, Liang Z, Xing Y, Huang W, Xue Z. Self-supervised multi-modal knowledge graph contrastive hashing for cross-modal search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(12): 13744–13753. DOI:10.1609/aaai.v38i12.29280.
- [9] Sun Y, Wang M, Ma Y. Semantic-alignment Transformer and adversary hashing for cross-modal retrieval. *Applied Intelligence*, 2024, 54(17): 7581–7602. DOI:10.1007/s10489-024-05501-2.
- [10] Han K, Liu Y, Wei R, Zhou K, Xu J, Long K. Supervised hierarchical online hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024, 20(4): 1–23. DOI:10.1145/3632527.
- [11] Li F, Wang B, Zhu L, Li J, Zhang Z, Chang X. Cross-domain transfer hashing for efficient cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(10): 9664–9677. DOI:10.1109/tcsvt.2024.3374791.
- [12] Liu X, Li J, Nie X, Zhang X, Wang S, Yin Y. Scalable unsupervised hashing via exploiting robust cross-modal consistency. *IEEE Transactions on Big Data*, 2024, 10(4): 514–527. DOI:10.1109/tbdata.2024.3350541.
- [13] Song G, Huang K, Su H, Song F, Yang M. Deep ranking distribution preserving hashing for robust multi-label cross-modal retrieval. *IEEE Transactions on Multimedia*, 2024, 26(1): 7027–7042. DOI:10.1109/tmm.2024.3358995.

- [14] Wang J, Zeng Z, Chen B, Wang Y, Liao D, Li G, Xia S T. Hugs bring double benefits: Unsupervised cross-modal hashing with multi-granularity aligned Transformers. *International Journal of Computer Vision*, 2024, 132(8): 2765–2797. DOI:10.1007/s11263-024-02009-7.
- [15] Li M, Ge M. Enhanced-similarity attention fusion for unsupervised cross-modal hashing retrieval. *Data Science and Engineering*, 2025, 10(2): 258–276. DOI:10.1007/s41019-024-00274-7.
- [16] Sun Y, Dai J, Ren Z, Chen Y, Peng D, Hu P. Dual self-paced cross-modal hashing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(14): 15184–15192. DOI:10.1609/aaai.v38i14.29441.
- [17] Zuo R, Zheng C, Li F, Zhu L, Zhang Z. Privacy-enhanced prototype-based federated cross-modal hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024, 20(9): 1–19. DOI:10.1145/3674507.
- [18] Tu J, Liu X, Hao Y, Hong R. A unified generative hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025, 21(12): 1–15. DOI:10.1145/3744567.
- [19] Adilakshmi K, Srinivas M, Kodali A, Srilakshmi V. Optimized RMDL with transfer learning for sentiment classification in the MapReduce framework. *Journal of Web Engineering*, 2023, 22(8): 1101–1132. DOI:10.13052/jwe1540-9589.2282.
- [20] Medhat S, Abdel-Galil H, Aboutabl A E, Saleh, H. Iterative magnitude pruning-based light-version of AlexNet for skin cancer classification. *Neural Computing and Applications*, 2024, 36(3): 1413–1428. DOI:10.1007/s00521-023-09111-w.
- [21] Li X. Design of a Web content personalized recommendation system based on collaborative filtering improved by combining k-means and LightGBM. *Journal of Web Engineering*, 2025, 24(2): 267–290. DOI:10.13052/jwe1540-9589.2425.
- [22] Hu P, Zhu H, Lin J, Peng D, Zhao Y P, Peng X. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3877–3889. DOI:10.1109/tpami.2022.3177356.
- [23] Zhu L, Wu X, Li J, Zhang Z, Guan W, Shen H T. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge & Data Engineering*, 2023, 35(09): 8838–8851. DOI:10.1109/tkde.2022.3218656.

- [24] Wu Q, Zhang Z, Liu Y, Zhang J, Nie L. Contrastive multi-bit collaborative learning for deep cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 2024, 36(11): 5835–5848. DOI:10.1109/TKDE.2024.3419577.

Biographies



Yang-hao Li (June 1996), male, holds a master's degree in Business Administration from Guangdong University of Technology, China. Currently working as an Assistant Engineer at China Southern Power Grid Artificial Intelligence Technology Co. Ltd. His research and work focus on large models for the power industry, power informatization and power marketing.



Zhao-jie Dong (September 1985), male, holds a master's degree in Computer Technology from Sun Yat-sen University, China. Currently, he works as a Senior Engineer at China Southern Power Grid Artificial Intelligence Technology Co. Ltd. His work and research focus on artificial intelligence, large language models and intelligent customer service.



Shi-song Wu (April 1986), male, holds a master's degree in Software Engineering from Tsinghua University, China. Currently, he works as a Senior Engineer at China Southern Power Grid Artificial Intelligence Technology Co. Ltd. His research focuses on artificial intelligence, large language models and intelligent customer service.



Xuan-ang Li (June 1993), male, holds a master's degree in Computer Technology from Guangxi University, China. Currently, he works as an Engineer at China Southern Power Grid Artificial Intelligence Technology Co. Ltd. His research focuses on artificial intelligence and power informatization.



Lian-yu Sha (November 1990), male, obtained his bachelor's degree in Computer Technology from Harbin Normal University, China. Currently, he works as an Engineer at China Southern Power Grid Artificial Intelligence Technology Co. Ltd. His research and work focus on artificial intelligence, data analysis and power marketing.