Journal of Engineering, Project, and Production Management
2026, 16(6), 025-185

A Deep Feature Fusion Framework for Accurate and Efficient Cross-Modal Retrieval in Digital Libraries

Xiaoyu Sun[1], Wenjie Meng[2], and Xuesong Zhang[3]
[1] Associate Researcher, Library of the China University of Petroleum (East China), Qingdao, 266580, China, E-mail: sunxiaoyu0818@126.com (corresponding author).
[2] Librarian, Library of the China University of Petroleum (East China), Qingdao, 266580, China
[3] Associate Researcher, Library of the China University of Petroleum (East China), Qingdao, 266580, China

Project Management
Received August 27, 2025; revised December 21, 2025; June 5, 2026; accepted June 9, 2026
Available online June 17, 2026

Abstract: To address the multi-modal resource retrieval needs of digital libraries, archives, and knowledge bases, this study proposes a Feature Fusion Cross Modal Hashing (FFCMH) model. It innovatively constructs a specialized dataset (multi-level filtering + cross modal denoising), employs autoencoders for feature fusion, and enhances image-text extraction through semantic segmentation, thereby supporting efficient retrieval. Experimental findings reveal that the proposed technology outperforms existing mainstream models, such as Locality-Sensitive Hashing, Semantic Topic Multi-modal Hashing (STMH), and Deep Cross Modal Hashing (DCMH), across metrics including recall rate and average precision on professional datasets. For instance, on the Flickr-25k dataset, the proposed technology achieves a maximum recall rate of 95.5% for Image-To-Text Cross Modal (I2TCM) retrieval and 86.7% for Text-To-Image Cross Modal (T2ICM) retrieval. Furthermore, the proposed technology exhibits significant advantages in retrieval accuracy and efficiency on self-made datasets, with average precision values of 0.948 for I2TCM and 0.938 for T2ICM, while requiring significantly less retrieval time than other models. This technology provides technical support for libraries to achieve efficient and precise cross modal resource retrieval.

Keywords: Library management, cross-modal hashing, feature fusion, semantic segmentation, resource retrieval.

DOI 10.32738/JEPPM-2025-185

1. Introduction

In the digital age, library book resource management encounters numerous difficulties. As book collection continues to grow steadily, traditional manual management approaches cannot meet the requirements for efficient, accurate management (Senthil et al., 2023). The frequent occurrence of problems such as book misplacement, loss, and damage, as well as the urgent need to address copyright protection, content quality testing, and integration management with paper books for electronic book resources (Barsha and Munshi, 2024). To enhance the effectiveness and standard of service for book resource management and meet the personalized needs of readers, it is particularly important to introduce advanced detection technology (Shamsitdinova et al., 2024). Wu et al. (2023) investigated user topic privacy leakage in Chinese book resource retrieval and proposed a client-side book search privacy protection framework. The framework employs an algorithm that modifies user query sequences, generating enhanced sequences through keyword substitution, removal, and insertion. The results showed that it could improve privacy security without affecting retrieval performance. Raghavendra et al. (2023) conducted research on the problem of low accuracy in algorithmic content retrieval from book resources. They improved the Seer search engine's algorithm and used machine learning methods to automatically identify relevant phrases and extract feature groups of attention content and structure. Tests showed that this technology significantly improved book retrieval effectiveness, outperforming similar technologies. Huang (2024) conducted research on the problem of insufficient library book resources, studied various intelligent retrieval systems, and tested the stability and accuracy of book retrieval algorithms. The results indicated that adopting intelligent book retrieval could enhance the reading experience of books and significantly improve the retrieval efficiency of traditional books.

Cross-Modal (CM) detection technology has important applications in resource query and retrieval. It achieves accurate book recognition and retrieval while improving the user experience by integrating multiple modalities, including images and text. Wang et al. (2023) investigated security issues in Deep Cross Modal Hashing (DCMH), retrieval and proposed a security detection method. The method extracted fine-grained target semantics through fusion modules and enhanced

attack performance by combining adversarial training discriminators. The results showed that the method had secure retrieval performance, ensuring data integrity and security. Li et al. (2024) studied the problem of poor retrieval performance in unsupervised CM hashing, which arises from the lack of semantic supervision. They proposed cross domain transmission hashing, used auxiliary domain reinforcement learning to construct auxiliary and target weak semantic correlation graphs, and designed a target CM hashing network. The results showed advantages in retrieval accuracy and training efficiency, but insufficient adaptability to dynamically updated multi-source auxiliary domains. Huo et al. (2023) identified problems with existing deep CM hashing in multi-label retrieval and proposed a deep semantic-aware proxy hashing framework that captures multi-label correlations via a semantic-aware proxy loss. Tests showed that this technology had good stability and accuracy, but its efficiency was relatively low in large-scale testing. Meng et al. (2023) studied the impact of modal private information on semantic embedding in CM hashing, proposed semantic decoupling adversarial hashing, decoupled modal public and private features, and introduced a variational information bottleneck. Tests demonstrated that this technology could substantially enhance the retrieval performance of CM information.

According to the above research, as library digitization accelerates, book resource retrieval technology has advanced in privacy protection, algorithm optimization, intelligent applications, and other areas, thereby supporting resource management and querying (Hu et al., 2023). However, current technology has shortcomings, such as traditional techniques that mainly focus on querying textual information and neglect images and covers (Tu et al., 2023). At present, CM retrieval has efficient data processing and retrieval capabilities by integrating multi-modal information. Therefore, to address the problems faced by existing book resource retrieval models, such as poor CM retrieval performance, weak semantic correlation, and limited adaptability to complex images, a Feature Fusion Cross Modal Hashing (FFCMH) retrieval model is designed to support CM and enhance the effectiveness of library resource management. There are two innovations in the research. One is to construct a specialized multimodal book dataset, which addresses the poor adaptation of existing datasets by leveraging multi-level screening and CM denoising. The second is to consider feature fusion and propose a CM hash retrieval model that utilizes autoencoders to achieve CM feature fusion. Additionally, semantic segmentation is considered to enhance the extraction of graphic and textual information, thereby improving the effectiveness of book retrieval. This study can provide technical support for libraries to achieve efficient, accurate CM resource retrieval and to improve management and service levels.

2. Methods and Materials

2.1. Construction and Processing of Multi-Modal Book Dataset

The multi-modal resource retrieval of digital libraries faces problems such as modal heterogeneity, semantic noise, and uneven data quality, and the adaptability of existing general datasets is insufficient. To this end, the research constructs a specialized dataset that provides a foundation for subsequent library resource retrieval via multi-level screening and CM denoising. The construction and processing flow of the book dataset is shown in Fig. 1.

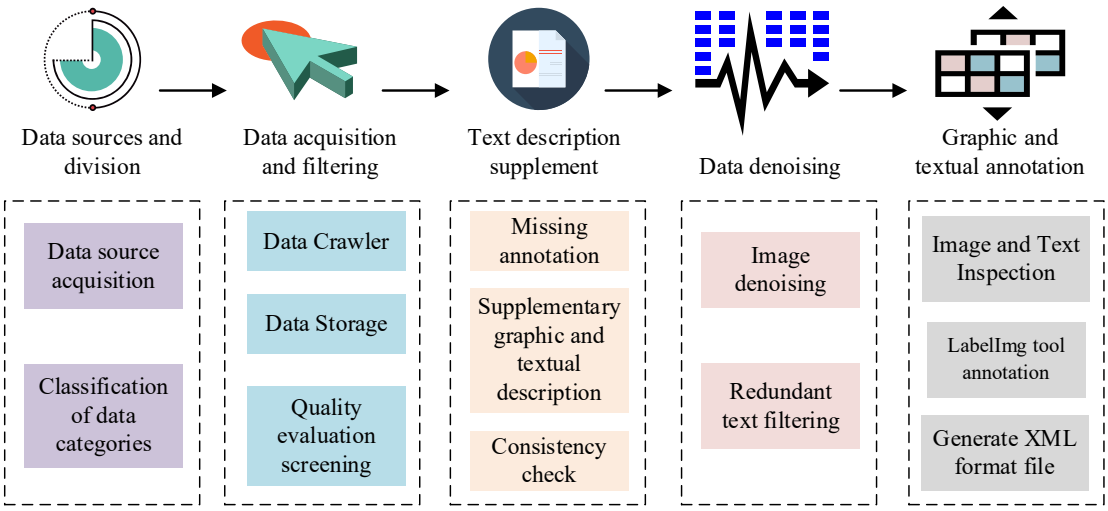

Fig. 1. Multi-modal graph dataset modeling technology process

According to Fig. 1, the technical process comprises five stages. The data source and category classification, data acquisition and filtering, text description supplementation, graphic and textual annotation, and noise reduction. Regarding data sources and category classification, reference is made to the relevant book content in the National Digital Library. The data comprises nine types of book theme resources, including intangible cultural heritage, multi-ethnic graphics and text, and new year paintings. The research on data acquisition uses targeted crawling, recording the URL and other associated metadata for each image as it crawls (Li et al., 2024). To ensure data quality, the study adopts an image quality screening mechanism, as shown in Eq. (1).

$$Q(I) = \alpha \cdot \text{Edge}(I) + (1-\alpha) \cdot \text{Contrast}(I) \tag{1}$$

In Eq. (1), $Q(I)$ is the quality score of the image, $\text{Edge}(I)$ represents the edge strength extracted by the Canny edge detection operator, $\text{Contrast}(I)$ is the image contrast, and α is the adjustment coefficient. When $Q(I) < \theta$ ($\theta = 0.3$ It is

judged as a blurred image and removed. Due to missing or abbreviated text annotations in the original image, a "multi-person collaborative annotation+consistency verification" mechanism is adopted to supplement the text description, as shown in Fig. 2.

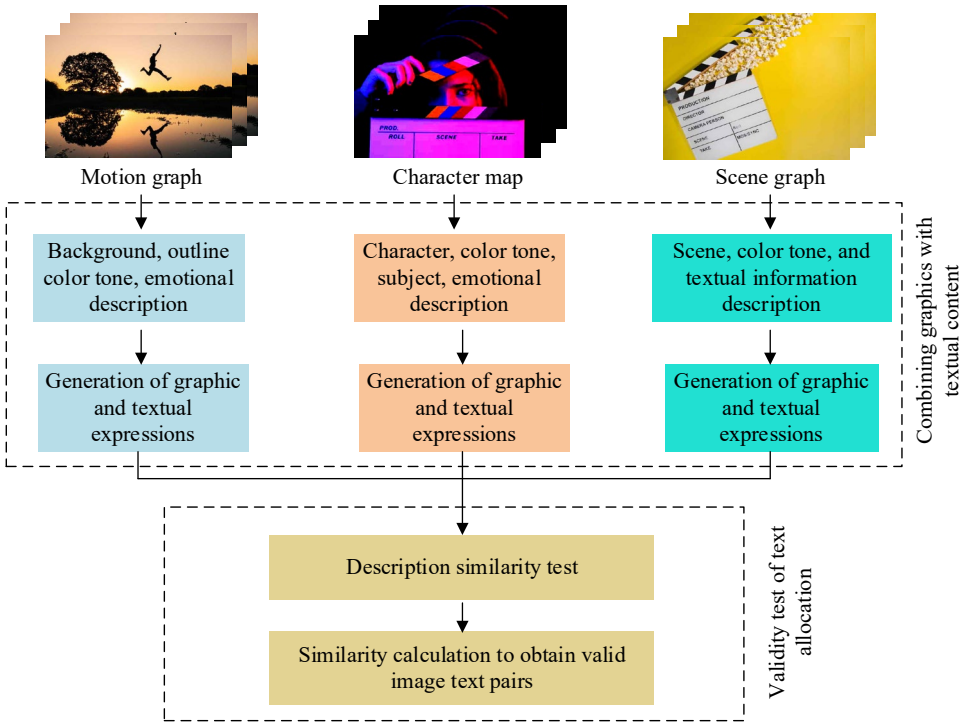

Fig. 2. Supplementary process for text and image description

According to Fig. 3, each image is independently generated by three annotators, each providing four descriptive texts that cover subjects, actions, scenes, and emotions in the image. The study uses a text similarity formula to verify consistency in descriptions, as shown in Eq. (2).

$$S(T_1,T_2)=\frac{\sum_{w\in T_1\cap T_2}\text{TF-IDF}(w)}{\sqrt{\sum_{w\in T_1}\text{TF-IDF}(w)^2}\cdot\sqrt{\sum_{w\in T_2}\text{TF-IDF}(w)^2}} \tag{2}$$

In Eq. (2), $S(T_1,T_2)$ is the cosine similarity between two texts T_1 and T_2, $\text{TF-IDF}(w)$ represents the weight of the word W. When the average similarity of three annotations <0.5 Once it is reached, a senior annotator re-annotates them to form a graphic-text pair, with 5 valid sentences corresponding to each image. In addition, the LabelImg tool is used for graphic and textual annotation, selecting key areas in the image with rectangular boxes and assigning labels. Additionally, corresponding text descriptions are associated to generate an Extensible Markup Language format annotation file containing the image path, label coordinates, and text content. To address data noise, denoising is necessary. Optimize low-quality images using the non-local mean denoising algorithm, as shown in Eq. (3).

$$I_{\text{denoised}}(x)=\frac{1}{C(x)}\sum_{y\in\Omega(x)}\exp\left(-\frac{\|I(x)-I(y)\|^2}{h^2}\right)\cdot I(y) \tag{3}$$

In Eq. (3), $I_{\text{denoised}}(x)$ is the grayscale value of the denoised image at the pixel x, $\Omega(x)$ is the neighborhood centered on x, h is the smoothing parameter ($h=10$), and $C(x)$ is the normalization coefficient. For text denoising, the study filters redundant text through word vector similarity and calculates the semantic matching degree between text and image labels as shown in Eq. (4).

$$M(T,L)=\frac{1}{|L|}\sum_{l\in L}\max_{t\in T}\text{Sim}(\text{vec}(t),\text{vec}(l)) \tag{4}$$

In Eq. (4), $M(T,L)$ represents the degree of text label matching, with a value range of [0,1], used to filter invalid text. L is the image label set, $\text{vec}(t)$ represents the Word to Vector (Word2Vec) of word t, and $\text{Sim}(\cdot)$ is the cosine similarity. When $M(T,L)<0.4$, the text is deemed invalid and removed.

2.2. Modeling of CM Hash Retrieval Based on Feature Fusion

After completing the construction of a multi-modal book dataset, a Feature Fusion-Based CM Hash Retrieval Model (FFCMH) is proposed to achieve efficient CM retrieval of text and image resources. The entire CM hash retrieval technology framework is shown in Fig. 3.

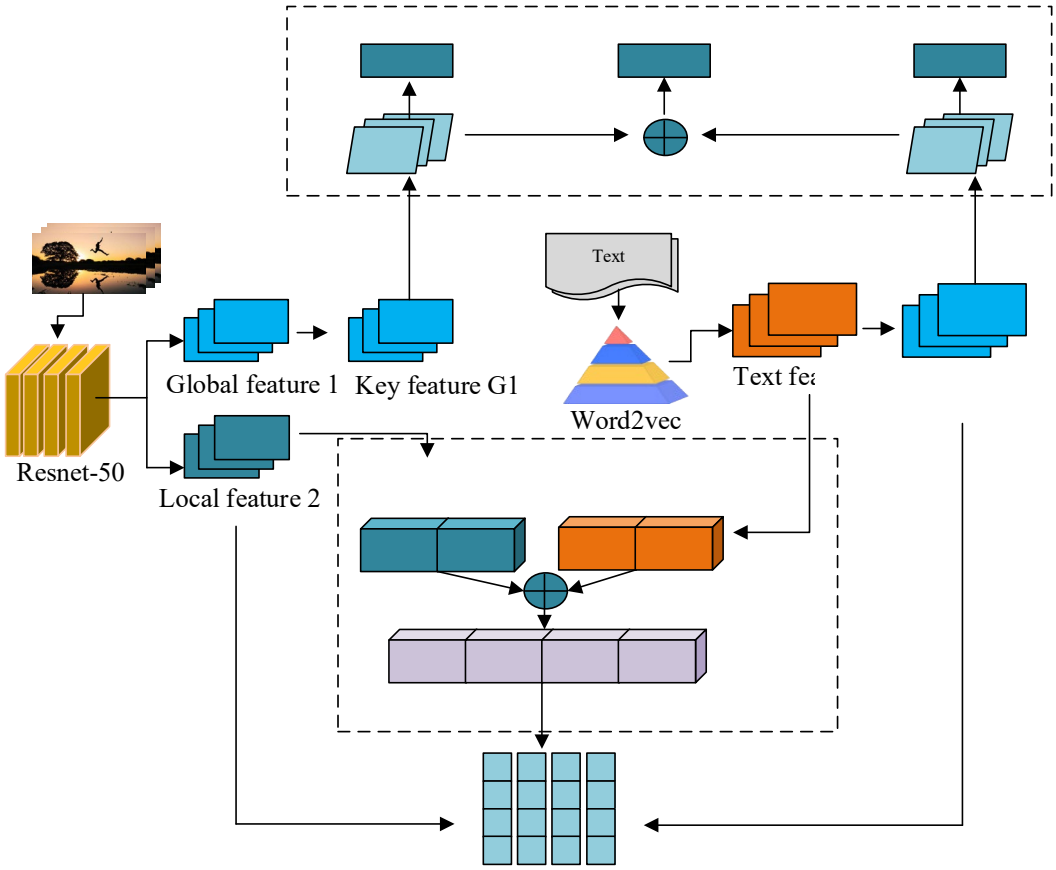

Fig. 3. CM hash retrieval technology framework

According to Fig. 3, the technical process includes four core steps: feature extraction, feature fusion, similarity matrix optimization, and hash code generation. Feature extraction is the foundation of CM retrieval, and the model must extract feature vectors with strong semantic representation from text and images separately. The study used the continuous bag-of-words algorithm in the Word2Vec model to convert the textual descriptions of book resources into dense, low-dimensional vectors (Wang et al., 2023). For the input text sequence $X = \{x_1, x_2, ..., x_n\}$ Contextual information is obtained by sliding a window, the probability of the central word is predicted, and the word vector space is iteratively optimized. The final generated text feature vector is shown in Eq. (5).

$$F = f(x_i; \theta_x) \tag{5}$$

In Eq. (5), θ_x is the parameter of the text feature extraction network. In addition, the study used Residual Network-50 (ResNet-50) to extract global and local features of images. For image data such as book covers and ancient book illustrations, $Y = \{y_1, y_2, ..., y_n\}$ the gradient vanishing problem is alleviated by connecting residual blocks across layers. After convolution and pooling operations, the global and local features $G = g(y_i; \theta_y)$ are output, where θ_y and θ_z are the parameters of the global and local feature extraction networks (Han et al., 2024). To compensate for the semantic differences between text and image features, an autoencoder was included in the original model to achieve CM fusion. The specific process is to input the text feature F and local image feature H into two encoders respectively, convert them into dimensional low dimensional features $encoder_1$ and $encoder_2$, and fuse the low dimensional features through an addition operation to obtain the fused feature $Encoder = encoder_1 + encoder_2$. In addition, the fused features are input into two decoders to restore the reconstructed features $\tilde{F}$ and $\tilde{H}$ in the same dimension as the original features, with the goal of minimizing the reconstruction error. The feature loss function is shown in Eq. (6).

$$O_{ff} = \left(\frac{1}{M} \sum_{i=1}^{M} \| x_i - \tilde{x}_i \|^2 + \frac{1}{M} \sum_{i=1}^{M} \| z_i - \tilde{z}_i \|^2 \right) / 2 \tag{6}$$

In Eq. (6), x_i and z_i are the local features of the input text and image, $\tilde{x}_i$ and $\tilde{z}_i$ are the corresponding reconstructed

features, and M is the number of samples. In addition, to improve the correlation between text semantics and image feature semantics, a multi-scale similarity matrix was designed in the model, including inter modal cosine similarity, neighborhood similarity, intra modal similarity, and total similarity. The cosine distance between the integrated text and image features is given in Eq. (7) (Fan et al., 2023).

$$S_d(x_i, y_j) = (1-\alpha)\cos(x_i, y_j) + \alpha\cos(x_i, y_j) \tag{7}$$

In Eq. (7), $\alpha \in [0,1]$ is the adjustment parameter used to balance the weights of the two modal information. For neighborhood similarity between modalities, the second-order similarity calculation using K-nearest-neighbor aggregated samples is studied, as shown in Eq. (8).

$$P(\sigma(x_i, y_j)) = \left[q \in N_k(i)\right] S_d(x_i, y_j) \tag{8}$$

In Eq. (8), $\sigma(x, y)$ denotes a predicate similar to x and y, $N_k(i)$ denotes the K nearest neighbor set of x_i, and $[\square]$ denotes Iverson brackets. For intra-modal similarity, the study expresses it through the cosine similarity of the fused features before and after fusion, as shown in Eq. (9) (Wu et al., 2024).

$$S_{dd} = \left(\frac{x_i \cdot \tilde{x}_i}{\| x_i \| \times \| \tilde{x}_i \|}\right) + \left(\frac{y_i \cdot \tilde{y}_i}{\| y_i \| \times \| \tilde{y}_i \|}\right) \tag{9}$$

In Eq. (9), $\tilde{x}_i$ and $\tilde{y}_i$ are the fused text and image features. Based on the above similarity, the matrix is dynamically updated as shown in Eq. (10).

$$\begin{cases} S(x_i, y_j) = (1-\beta)S_d + \beta S_{nn} + \alpha S_{ff} + S_{dd} + S_{ffh} \\ S_{t+1} = \kappa S_t + (1-\kappa)\tilde{S} \end{cases} \tag{10}$$

In Eq. (10), β is the adjustment parameter, $\tilde{S}$ is the newly learned similarity matrix, and $\kappa \in [0,1]$ is the update coefficient. By iteratively updating the similarity matrix, the model can continuously strengthen the semantic correlation between modalities. To achieve efficient retrieval, the model maps the fused features into binary hash codes and optimizes the semantic consistency of these codes using a loss function. The study uses symbolic functions to convert text features, image features, and their fusion into hash codes, as shown in Eq. (11).

$$\begin{cases} Hc_t = \text{sign}(F_i) \\ Hc_f = \text{sign}(G_i) \\ Hc = \text{sign}(F_i + G_i) \end{cases} \tag{11}$$

In Eq. (11), $\text{sign}(x) = 1$, if $x > 0$, otherwise it is -1. Hc is the hash code. Hc_t is the text hash code. Hc_f is the image hash code. The hash loss function is defined to minimize the difference between hash codes, as shown in Eq. (12).

$$S_{ffh} = \| Hc - Hc_t \| + \| Hc - Hc_f \| \tag{12}$$

This function ensures the semantic consistency of different modal hash codes by constraining the distance between unified hash code Hc, text hash code Hc_t and image hash code Hc_f.

2.3. Modeling of CM Hash Retrieval Considering Semantic Segmentation

Although the CM hash retrieval model based on feature fusion has achieved effective matching between text and images, it still has two limitations when dealing with complex images in library resources, such as ancient book illustrations and multi-element covers. Firstly, image features are susceptible to background noise interference, and secondly, static text features are difficult to capture contextual semantic associations. To this end, based on the original Feature Fusion-Based CM Hash Retrieval Model (FFCMH) model, a Semantic Segmentation-based Enhanced CM Hashing Retrieval Model (SSDCMH+) is proposed. The entire technical process is shown in Fig. 4.

According to Fig. 4, the study added DeepLabv3+ to the original model to achieve high-precision graph-semantic segmentation and replaced Word2Vec with BERT to extract dynamic text features, thereby further improving retrieval robustness. The overall model process comprises four core steps: image semantic segmentation and feature purification, dynamic text feature encoding, CM hash mapping, and multi-objective loss optimization. Through fine feature extraction and deep semantic association, efficient retrieval of library resources is achieved. To accurately extract semantic regions from images, the model uses the DeepLabv3+ network for pixel-level segmentation and optimizes its backbone network design. It enhances image information extraction by introducing an Atrous Spatial Pyramid Pooling (ASPP) module and a decoder boundary repair mechanism (Wang et al., 2023). The entire improved DeepLabv3 network structure is shown in Fig. 5.

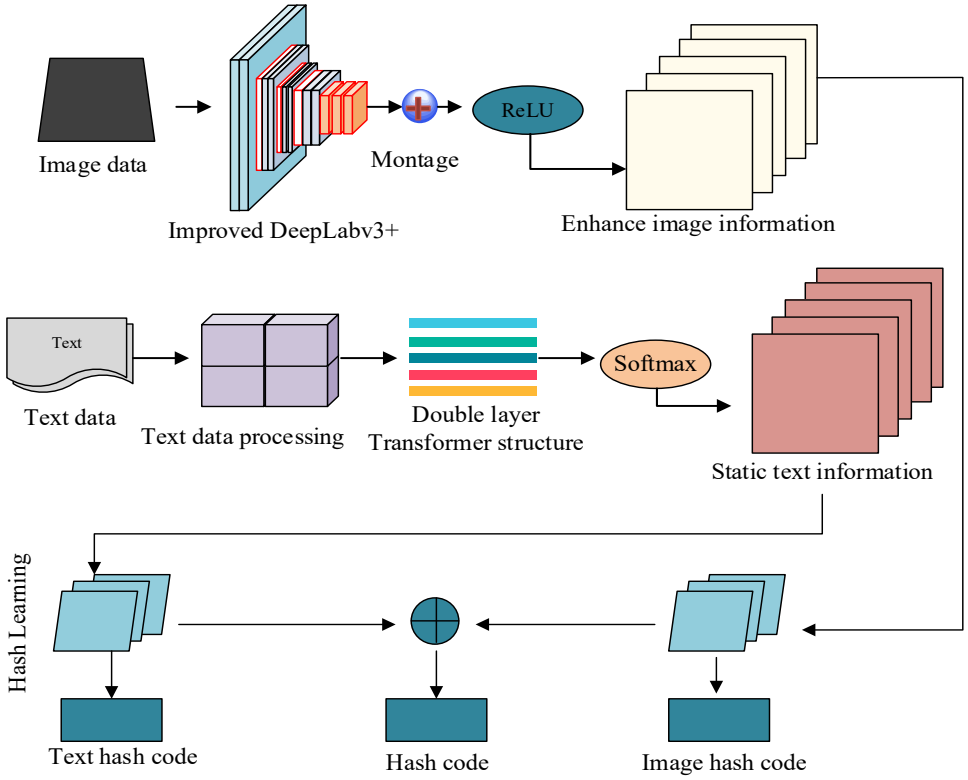

Fig. 4. Framework of an enhanced CM hash retrieval model based on semantic segmentation

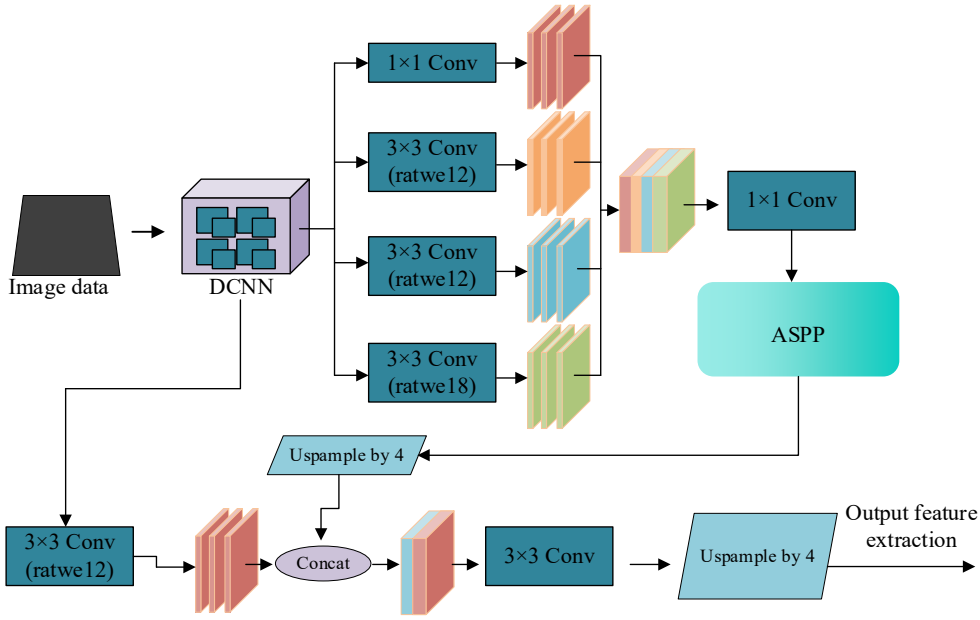

Fig. 5 Improved DeepLabv3+ network architecture

According to Fig. 5, the backbone network uses an improved version of Xception as the feature-extraction backbone and replaces the traditional pooling layer with a separable convolutional layer with holes. This structure decomposes the standard convolution into a Depthwise Convolution Neural Network (DCNN) and a 1×1 pointwise convolution, reducing the number of convolutional parameters while increasing the control receptive field size. In addition, an ASPP module is introduced at the end of the network encoder to achieve multi-scale feature fusion. Among them, dilated convolution expands the receptive field by inserting zeros into the convolutional kernel, while global pooling captures image-level context. In addition, the study introduces a boundary repair mechanism for the decoder, which fuses the low resolution feature map output by ASPP with the high-resolution shallow features generated by the backbone network. After passing through a 1×1 convolutional compression channel, the shallow features are added pixel-by-pixel to the upsampled deep features, and finally, a semantic segmentation map of the same size as the input image is generated via 4× upsampling. In text feature encoding, to overcome the limitations of static word vectors in Word2Vec, the BERT-base model is used to extract text features, with its core improvement being the capture of contextual dependencies via bidirectional Transformers

(Yang et al., 2023). Firstly, in the word segmentation process, the position encoding generated by the sine function is expressed as Eq. (13).

$$PE_{(pos,2i)} = \sin\left(pos / 10000^{2i/d} \right) \tag{13}$$

In Eq. (13), pos is the position index of the word in the sequence, i is the vector dimension index, and $d = 768$ is the BERT hidden layer dimension. In addition, the encoding utilizes the periodicity of trigonometric functions to enable the model to distinguish differences in word order, as shown in Eq. (14).

$$PE_{(pos,2i+1)} = \cos\left(pos / 10000^{2i/d} \right) \tag{14}$$

In Eq. (14), $2i+1$ represents the odd dimensional index, which forms orthogonal encoding with the sine function of even dimensions to enhance the uniqueness of positional information. In the encoder part, a twelve-layer bidirectional Transformer structure is used, and self-attention calculates the correlation weights between words by scaling dot products as shown in Eq. (15) (Jiang et al., 2024).

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \right) V \tag{15}$$

In Eq. (15), Q is the query matrix, K is the key matrix, and V is the value matrix. QK^T is the similarity matrix, which is divided by $\sqrt{d_k}$ to alleviate gradient vanishing. The softmax function normalizes the weights to the [0,1] interval. This mechanism can capture the semantic association between "library" and "collection resources" and generate a 768-dimensional vector as text feature F_t . The model maps the segmented image features and BERT text features to a unified hash space. For the hash code mapping of image feature F_i , a hash code is generated through a sign function as shown in Eq. (16) (Preethi and Mamatha, 2023).

$$H_i = \text{sign}(F_i) \tag{16}$$

In Eq. (16), $H_i \in \{-1,1\}^{64}$ is the 64-bit image hash code and $\text{sign}(x)$ is the sign function. For the hash code mapping of text feature F_t , the study directly binarizes the output F_t of BERT, as shown in Eq. (17).

$$H_t = \text{sign}(F_t) \tag{17}$$

In Eq. (17), $H_t \in \{-1,1\}^{64}$ is a 64-bit text hash code. In this study, the CM unified hash code $H = \text{sign}(W_F F_i + W_T F_t)$ is generated by feature fusion, in which W_F and W are modal weight matrices, which are used to balance the contribution of image and text features, and dynamically adjusted by training. Finally, the research uses the improved cross-entropy loss to measure semantic matching, as shown in Eq. (18).

$$\mathsf{L}_1 = -\sum_{m,n=1}^{N} S_{mn} \log\left(\sigma(\text{ƛ}_{mn})\right) + (1 - S_{mn}) \log\left(1 - \sigma(\text{ƛ}_{mn})\right) \tag{18}$$

In Eq. (18), $S_{mn} \in \{0,1\}$ is the semantic label, N is the total sample size, $\text{ƛ}_{mn} = F_i^T F_t / \sqrt{768}$ is the feature space similarity, σ is the Sigmoid function, and maps the similarity to the [0,1] interval.

3. Results

3.1. CM Hash Retrieval Experiment Based on Feature Fusion

To verify the effectiveness of the proposed technology in CM book resource retrieval, it was compared with existing mainstream models. The experimental environment parameters are shown in Table 1.

Table 1. Experimental environment

Project	Parameter
Graphics card	NVIDIA GeForce RTX 3060
Processor	I7 12700K
Storage	1TB SSD
Memory	64GB DDR5
Operating system	Ubuntu 22.04 LTS
Learning framework	PyTorch 2.0
Programming software	Python 3.9

According to Table. 1, the operating system was Ubuntu 22.04LTS, and the deep learning framework used PyTorch 2.0. The study utilized its Torch compile feature to accelerate model training. The experimental data included the public Flickr-25k dataset and the National University of Singapore-wide (NUS-wide) dataset. The Flickr-25k dataset contains 25000 images and corresponding text descriptions, covering various daily scenarios, and is a commonly used benchmark dataset for CM retrieval. The NUS-wide dataset contains 196843 images, associated with 81 concept labels and text descriptions. It has a large scale and rich scenes, rendering it appropriate for testing large-scale data performance. In addition, the study established a self-developed dataset covering nine types of book-themed resources, including multi-ethnic graphic and textual information, multicultural heritage and ancient dust and shadows. After screening and noise reduction processing, it contained 10800 samples, each with an image corresponding to a bibliographic description, for targeted testing of library resource retrieval performance. The study introduced Locality Sensitive Hashing (LSH), Semantic Topic Multi-modal Hashing (STMH), and Deep CM Hashing (DCMH) as comparative models. The study selected Flickr-25k as the basis and compared the recall retrieval performance of different techniques, as shown in Fig. 6.

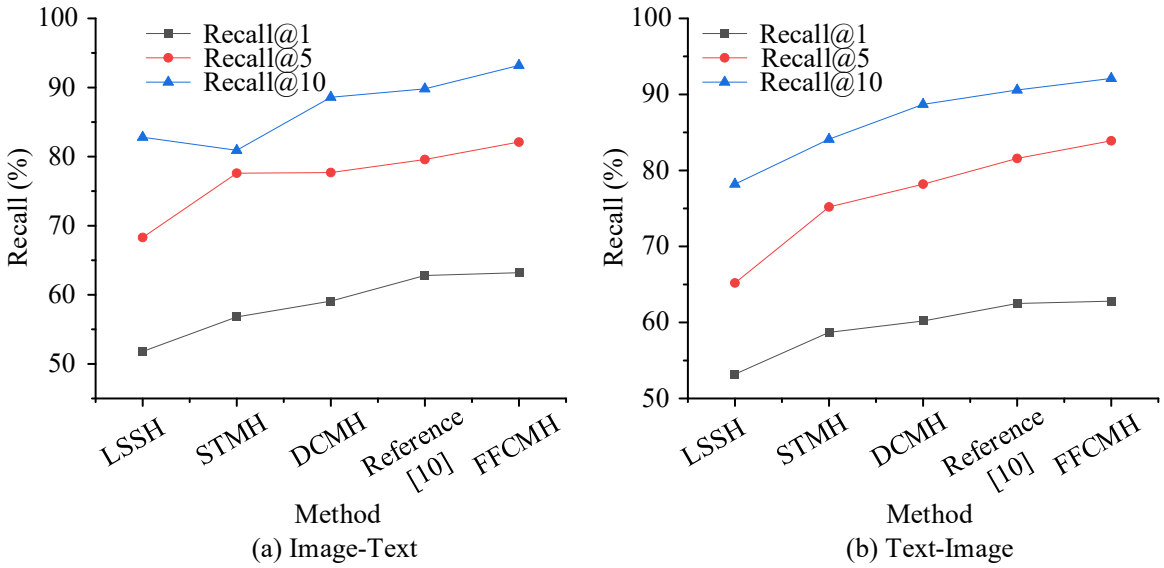

Fig. 6. Comparison of recall rates in CM retrieval of Flickr-25k dataset

Fig. 6(a) presents the results of Image-To-Text Cross Modal (I2TCM) retrieval, revealing that the recall rate gradually increased as the number of recommendations rose. When the number of recommendations was set to ten, all five models achieved their optimal recall rates, with FFCMH, the model in Meng et al. (2023), DCMH, STMH, and Locality Sensitive Hashing (LSH), yielding recall values of 95.5%, 92.4%, 90.1%, 82.8%, and 84.7%, respectively. It was evident that the FFCMH model proposed by the research performed the best. The main reason was that the research model used autoencoders to deeply fuse 4096-dimensional text and 2048-dimensional image features, eliminating modal heterogeneity. Fig. 6(b) shows the results of Text-To-Image Cross Modal (T2ICM) retrieval, where FFCMH exhibited the best performance, achieving a recall rate of 86.7% in the test with five recommendations, surpassing the 83.8% and 78.7% of the model in Meng et al. (2023) and DCMH, respectively. In contrast, LSH and STMH demonstrated relatively poor overall performance, with T2ICM retrieval recall rates significantly lower than those of I2TCM retrieval. The research tested the Mean Average Precision (MAP) of different techniques using the Flickr-25k dataset, as illustrated in Fig. 7.

Fig. 7(a) displays the MAP results for I2TCM retrieval. It was observed that FFCNH and the model in Meng et al. (2023) significantly outperformed other retrieval models, and the retrieval performance improved as the number of hash code bits increased. When the hash code length was 128 bits, the MAP values for FFCNH and the model in Meng et al. (2023) were 0.824 and 0.818, respectively, while LSH exhibited the poorest performance with an MAP value of 0.776. The research model used feature fusion to improve the semantic representation's accuracy, and as the number of hash code bits increased, the details of the fused features were better preserved. Fig. 7(b) presents the MAP results for T2ICM retrieval. LSH and STMH demonstrated notably inferior retrieval performance compared to I2TCM retrieval. For instance, when the hash code length was sixteen bits, the MAP values for LSH and STMH were 0.612 and 0.645, respectively, whereas DCMH (Meng et al., 2023) and FFCMH achieved MAP values of 0.895 and 0.723, respectively. It was evident that FFCMH exhibited the best stability and precision, performing optimally in both directions of CM retrieval. Subsequently, the research tested the image-text retrieval accuracy of different techniques using the NUS-wide dataset, as illustrated in Fig. 8.

Fig. 8(a) presents the accuracy results of text-based CM retrieval. As the scale of the retrieval samples increased, the retrieval accuracy of all five models declined. Overall, FFCMH demonstrated the best retrieval performance. Overall, FFCMH had the best retrieval performance, as the dynamic similarity matrix could iteratively optimize semantic associations with increasing data volume and resist data redundancy interference. For instance, when the retrieval sample size was 4,000, FFCMH achieved a retrieval accuracy of 83.8%, outperforming the models in Meng et al. (2023), DCMH, STMH, and LSH, which recorded accuracies of 81.2%, 80.2%, 78.2%, and 74.4%, respectively. Fig. 8(b) displays the accuracy results of T2ICM retrieval. In this scenario, FFCMH consistently maintained the highest retrieval accuracy and

stability, significantly outperforming the next best model in Meng et al. (2023). Its average retrieval accuracy reached 90.5%, while the models in Meng et al. (2023), DCMH, STMH, and LSH achieved accuracies of 84.7%, 80.2%, 78.4%, and 75.6%, respectively.

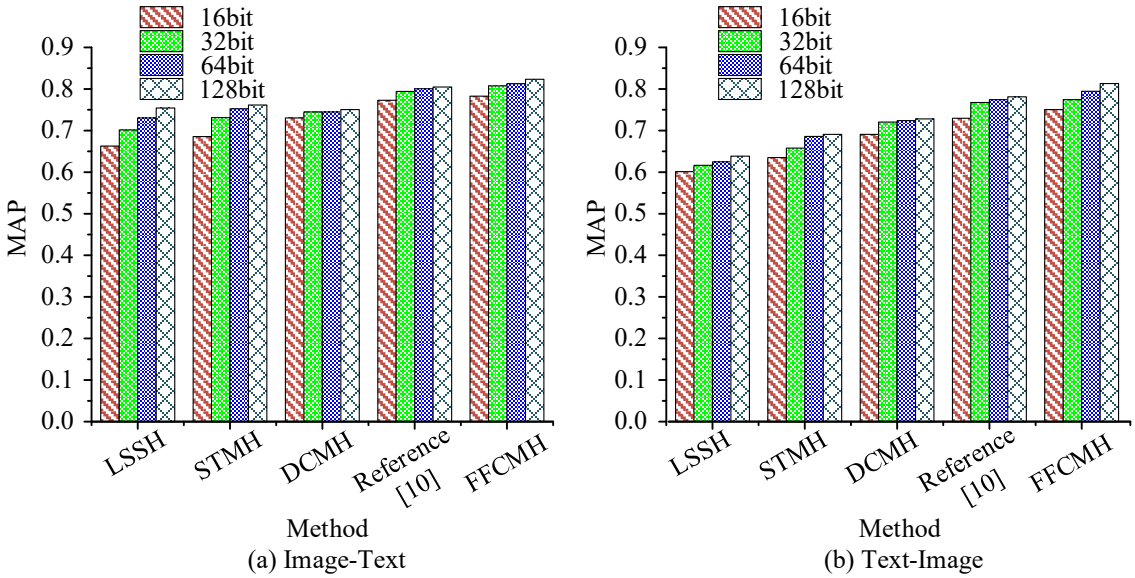

Fig. 7. Comparison of CM retrieval Mean Average Precision (MAP) in Flickr-25k dataset

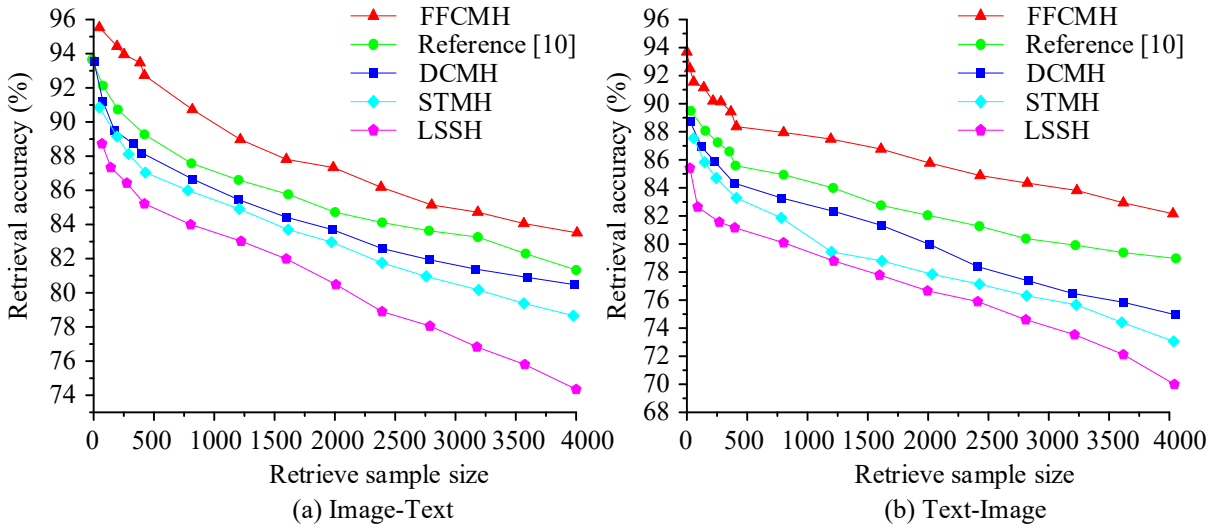

Fig. 8. Comparison of CM retrieval accuracy in NUS wide large scale dataset

3.2. CM Hash Retrieval Experiment Considering Semantic Segmentation

Considering the semantic differences between CM text and image retrieval, this will directly affect retrieval performance in complex scenes. Especially when there is significant noise between the image and text, the effectiveness of FFCMH image text retrieval is limited. Therefore, Wang et al. (2023), who introduced support for semantic segmentation technology, were compared with the proposed SSDCMH+. The retrieval MAP results of different techniques under self-made data are shown in Fig. 9.

Fig. 9(a) presents the I2TCM retrieval results on the self-made dataset. It was found that SSDCMH+ placed greater emphasis on the semantic correlation between images and text, demonstrating significant advantages over comparable techniques. For instance, when the hash code length was 128 bits, the MAP values for SSDCMH+, FFCMH, the models in Wang et al. (2023) and Yang et al. (2023), DCMH, and STMH were 0.948, 0.875, 0.863, 0.854, 0.761, and 0.724, respectively. Fig. 9(b) displays the T2ICM retrieval results on the self-made dataset. In this scenario, SSDCMH+ exhibited remarkable performance, achieving a maximum MAP of 0.938 with a 128-bit hash code, significantly outperforming the model in Wang et al. (2023), which reported a MAP of 0.864. A comparison of the retrieval time consumption of different models on the self-made dataset is shown in Fig. 10.

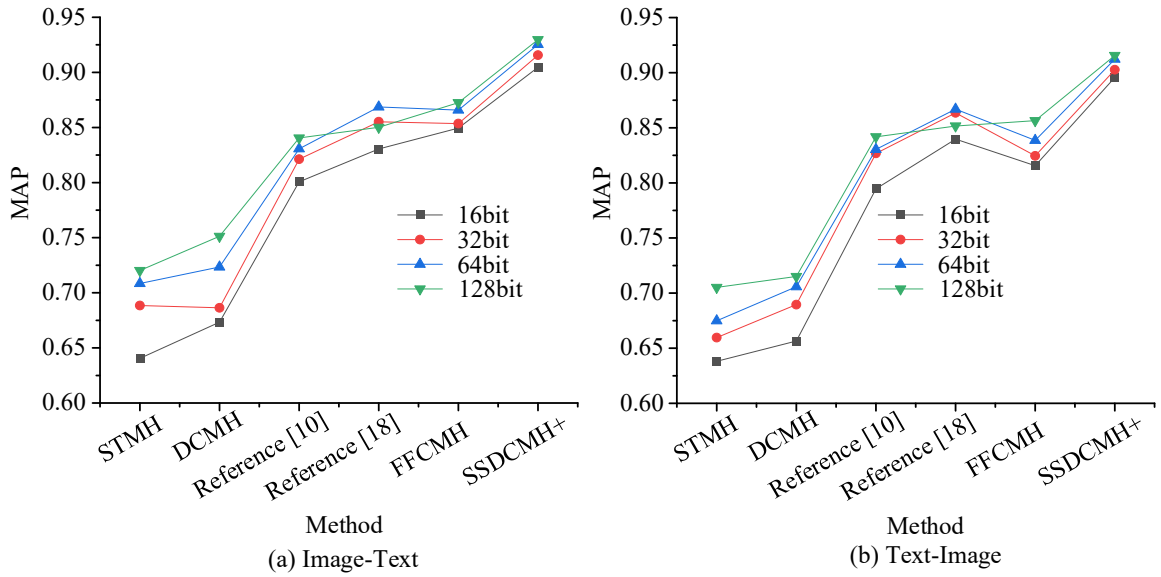

Fig. 9. Comparison of CM retrieval MAP in self-made library datasets

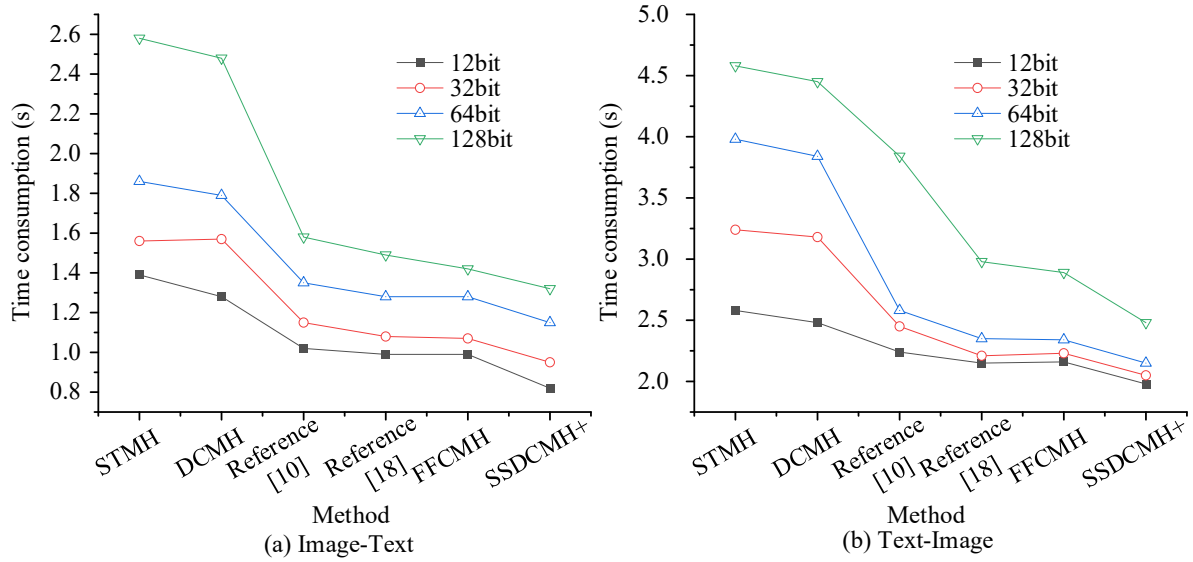

Fig. 10. Retrieval time test on self made dataset

Fig. 10(a) presents the time consumption results for I2TCM retrieval. The STMH model, lacking CM semantic processing capabilities, exhibited the longest retrieval time at 2.60 seconds with a 128-bit hash code. In contrast, SSDCMH+ demonstrated significant advantages, with a retrieval time of 1.38 seconds, while FFCMH and the model in Wang et al. (2023) recorded times of 14.25 seconds and 15.12 seconds, respectively. The main reason was that the research model adopted semantic segmentation to extract the effective image features in advance, reducing redundant computation time for text-to-image CM during subsequent hash mapping. The time consumption for T2ICM retrieval is shown in Fig. 10(b). In this scenario, when the hash code length was 128 bits, SSDCMH+ displayed an absolute advantage over other models, with a retrieval time of only 2.54 seconds, outperforming FFCMH's 3.05 seconds and significantly better than STMH's 4.68 seconds. Finally, the study introduced precision-recall curves to compare the testing effects of different techniques on a self-made dataset for image-text supplementation, as illustrated in Fig. 11.

Fig. 11(a) presents the test results on the self-made dataset without image-text supplementation. In this scenario, SSDCMH+ had the largest Area Under the Curve (AUC), indicating the best performance, with an AUC of 0.954, while FFCMH had an AUC of 0.924. STMH performed the worst, with an AUC of only 0.681. Fig. 11(b) displays the test results on the self-made dataset with image-text supplementation. It was found that the retrieval performance of all six retrieval models improved significantly, demonstrating that image-text supplementation in the dataset could notably enhance book retrieval accuracy. For instance, the AUC value for SSDCMH+ increased to 0.967, while those for FFCMH and the model in Meng et al. (2023) increased to 0.935 and 0.922, respectively. However, STMH and DCMH lacked the ability to capture dynamic semantics, so the AUC improvement was smaller, confirming SSDCMH+'s semantic adaptation advantage. The proposed technique demonstrated favorable performance in CM book retrieval.

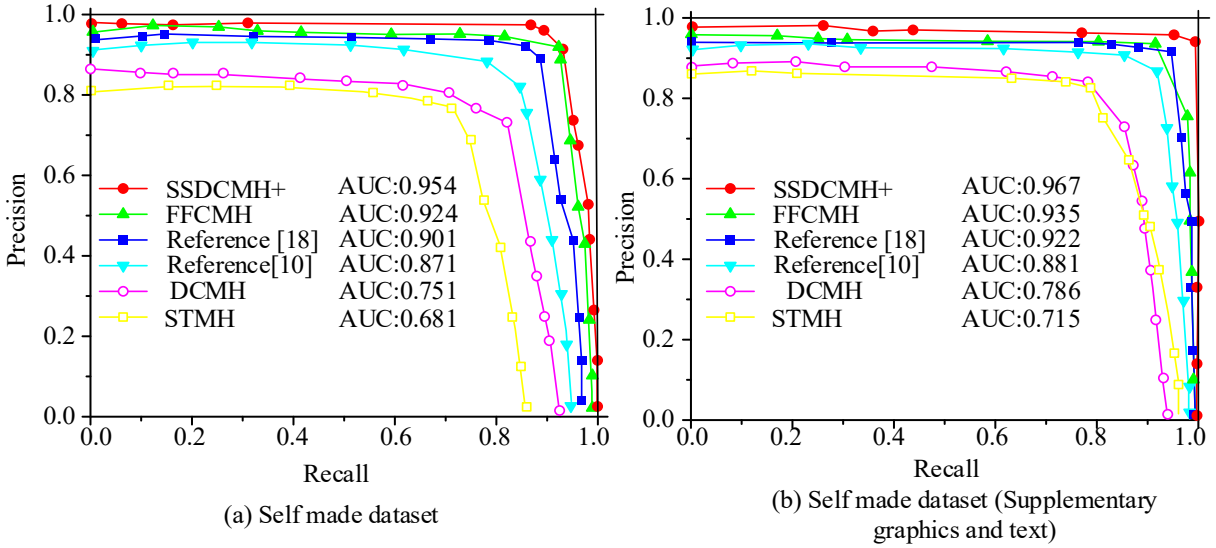

Fig. 11. Comparison of Accuracy Recall Ratio (AUC) in CM retrieval of multiple data sets

4. Conclusion

With the acceleration of library digitization, traditional retrieval methods are unable to meet the efficient matching requirements of multimodal resources and often suffer from low efficiency and poor accuracy in book retrieval. To address the aforementioned issues, a CM hash retrieval model based on feature fusion was proposed to improve the efficiency of library resource management.

Experimental tests demonstrated that FFCMH achieved a recall rate of 95.5% for I2TCM retrieval and 86.7% for T2ICM retrieval on the Flickr-25k dataset, significantly outperforming mainstream models such as LSH and STMH. Its advantage stemmed from the autoencoder's ability to deeply integrate 4096-dimensional text features with 2048-dimensional image features, while the dynamically updated similarity matrix strengthened inter-modal semantic associations, reducing retrieval deviations caused by CM differences. On the large-scale NUS-wide dataset, even with a sample size of 4,000, FFCMH maintained an I2TCM retrieval accuracy of 83.8%, indicating that the proposed technique was better at adapting to massive data. In scenarios emphasizing semantic associations between text and images, SSDCMH+ demonstrated outstanding performance on the self-made dataset, achieving an MAP of 0.948 for I2TCM retrieval and 0.938 for T2ICM retrieval, outperforming all comparable techniques. In terms of retrieval time for I2TCM tasks, SSDCMH+ required only 1.38 seconds, compared to 14.25 seconds for FFCMH and 15.12 seconds for the model in Wang et al. (2023). Additionally, in precision-recall testing, SSDCMH+ excelled, with an ROC value of 0.967 on the self-made dataset with image-text supplementation, representing the best performance. This was attributed to its use of DeepLabv3+ to enhance semantic segmentation of complex images and to adopting the BERT model instead of Word2Vec to capture textual contextual dependencies, thereby strengthening the analysis of associations across different semantic information and further improving retrieval effectiveness. In addition, the research expanded the data to a million-level database for testing and research, using data fusion of NUS-wide datasets, self-developed datasets, and public multimodal libraries. The proposed SSDCMH+ also showed excellent performance, similar to FFCMH and the literature (Wang et al., 2023), which encountered problems such as hash code retrieval and matching delays, and abnormal memory resource occupation. However, SSDCMH+ still performed best, with a CM retrieval time of only 2.73s, far lower than FFCMH and the literature (Wang et al., 2023) at 7.64s and 11.55s, and had the lowest resource occupation. In addition, the study tested multilingual scenarios using data from English, Japanese, and French collections. The cross language MAP of the SSDCMH+model was 0.913, showing the best overall performance and verifying its good adaptability across other language scenarios.

The technology proposed by the research showed strong practical effects. However, the research techniques also have shortcomings. The research mainly focused on searching Chinese and English data, without considering other text and multi-cover book information. Moreover, the deployment of technology must account for library scale issues, such as the need for more robust service systems and hardware for large libraries, while balancing cost and computational power. Additionally, IT platform integration required addressing compatibility with existing library collection management systems to avoid data format conflicts and reduce the time and errors associated with migrating historical multi-modal resources. Future efforts should focus on enhancing multi-source information analysis and processing capabilities, optimizing lightweighting to reduce hardware requirements, developing standardized integration interfaces, and promoting the practical application of technology across libraries of varying sizes.

Funding

The research is supported by the 2023 Qingdao Philosophy and Social Science Planning Research Project, Research on Collection Allocation Strategies for Multi-Campus University Libraries in Qingdao under the Digital-Intelligent Context (NO. QDSKL2301043).

Author Contributions

Xiaoyu Sun contributed to methodology, data curation and writing the original draft. Wenjie Meng contributed to conceptualization, software, formal analysis and writing the original draft. Xuesong Zhang contributed to conceptualization, methodology, software and formal analysis.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used DeepSeek only for language editing and formatting assistance. The authors reviewed and took full responsibility for all content.

References

Barsha, S. and Munshi, S. A. (2024). Implementing Artificial Intelligence in Library Services: A Review of Current Prospects and Challenges of Developing Countries. *Library Hi Tech News*, 41(1), 7-10. doi: https://doi.org/10.1108/LHTN-07-2023-0126.

Fan, W., Zhang, C., Li, H., Jia, X., and Wang, G. (2023). Three-Stage Semisupervised Cross-Modal Hashing with Pairwise Relations Exploitation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1), 260-273. doi: 10.1109/TNNLS.2023.3263221.

Han, L., Paoletti, M. E., Moreno-Álvarez, S., Haut, J. M. and Plaza, A. (2024). Deep Shared Proxy Construction Hashing for Cross-Modal Remote Sensing Image Fast Target Retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218(15), 44-56. doi: 10.1016/j.isprsjprs.2024.10.004.

Hu, P., Huang, Z., Peng, D., and Wang, X. (2023). Cross-Modal Retrieval with Partially Mismatched Pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 45(8), 9595-9610. doi: 10.1109/TPAMI.2023.3247939

Huang, Y. H. (2024). Exploring the Implementation of Artificial Intelligence Applications Among Academic Libraries in Taiwan. Library Hi Tech, 42(3), 885-905. doi: https://doi.org/10.1108/LHT-03-2022-0159.

Huo, Y., Qin, Q., Dai, J., Wang, L., Zhang, W., and Huang, L. (2023). Deep Semantic-Aware Proxy Hashing for Multi-Label Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1), 576-589.

Jiang, X., Fan, J., Zhang, J., Lin, Z., and Li, M. (2024). Multilevel Deep Semantic Feature Asymmetric Network for Cross-Modal Hashing Retrieval. *IEEE Latin America Transactions*, 22(8), 621-631. doi: 10.1109/TLA.2024.10620388.

Li, F., Wang, B., Zhu, L., Li, J., Zhang, Z., and Chang, X. (2024). Cross-Domain Transfer Hashing for Efficient Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology,* 34(10), 9664-9677. doi: 10.1109/TCSVT.2024.3374791.

Li, J., Wong, W. K., Jiang, L., Fang, X. Z., Xie, S., and Xu, Y. (2024). CKDH: CLIP-Based Knowledge Distillation Hashing for Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7), 6530-6541. doi: 10.1109/TCSVT.2024.3350695.

Meng, M., Sun, J., Liu, J., Yu, J., and Wu, J. (2023). Semantic Disentanglement Adversarial Hashing for Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3), 1914-1926. doi: 10.1109/TCSVT.2023.3293104.

Preethi, P. and Mamatha, H. R. (2023). Region-Based Convolutional Neural Network for Segmenting Text in Epigraphical Images//Artificial Intelligence and Applications, 1(2), 103-111. doi: 10.47852/bonviewAIA2202293.

Raghavendra Nayaka, P. and Ranjan, R. (2023). An Efficient Framework for Algorithmic Metadata Extraction Over Scholarly Documents Using Deep Neural Networks. *SN Computer Science*, 4(4), 341-351. doi: https://doi.org/10.1007/s42979-023-01776-3.

Senthil Kumaran, V. and Latha, R. (2023). Towards Personal Learning Environment by Enhancing Adaptive Access to Digital Library Using Ontology-Supported Collaborative Filtering. *Library Hi Tech*, 41(6), 1658-1675. doi: https://doi.org/10.1108/LHT-12-2021-0433.

Shamsitdinova, M., Khashimova, D., Niyazova, N., Nasirova, N., and Khikmatov, N. (2024). Harnessing AI for Enhanced Searching in Digital Libraries: Transforming Research Practices. *Indian Journal of Information Sources and Services*, 14(3), 102-109. doi: https://doi.org/10.51983/ijiss-2024.14.3.14.

Tu, R. C., Jiang, J., Lin, Q., Cai, C., Tian, S., and Wang, H. (2023). Unsupervised Cross-Modal Hashing with Modality-Interaction. *IEEE Transactions on Circuits and Systems for Video Technology,* 33(9), 5296-5308. doi: 10.1109/TCSVT.2023.3251395.

Wang, H., Zhao, K., and Zhao, D. (2023). A Triple Fusion Model for Cross-Modal Deep Hashing Retrieval. *Multimedia Systems,* 29(1), 347-359. doi: https://doi.org/10.1007/s00530-022-01005-6.

Wang, T., Zhu, L., Zhang, Z., and Shen, H. (2023). Targeted Adversarial Attack Against Deep Cross-Modal Hashing Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10), 6159-6172. doi: https://doi.org/10.48550/arXiv.2308.14263.

Wang, Y., Zhan, Y. W., Chen, Z. D., and Luo, X. (2023). Multiple Information Embedded Hashing for Large-Scale Cross-Modal Retrieval. I*EEE Transactions on Circuits and Systems for Video Technology*, 34(6), 5118-5131. doi: 10.1109/TCSVT.2023.3340102.

Wu, Q., Zhang, Z., Liu, Y., and Zhang, J. (2024). Contrastive Multi-Bit Collaborative Learning for Deep Cross-Modal Hashing. *IEEE Transactions on Knowledge and Data Engineering,* 36(11), 5835-5848. doi: 10.1109/TKDE.2024.3419577.

Wu, Z., Xie, J., Shen, S., Lin, C., Xu, G., and Chen, E. (2023). A Confusion Method for the Protection of User Topic Privacy in Chinese Keyword-Based Book Retrieval. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5), 2-19. doi: https://doi.org/10.1145/3571731.

Yang, F., Zhang, Q., Ding, X., Ma, F., Cao, J., and Tong, D. (2023). Semantic Preserving Asymmetric Discrete Hashing for Cross-Modal Retrieval. *Applied Intelligence*, 53(12), 15352-15371. doi: https://doi.org/10.1007/s10489-022-04282-w.

Xiaoyu Sun received her Master's degree in English from Toyama University in Japan in 2007. She is currently an Associate Researcher at the Library of the China University of Petroleum (East China), China. Her research focuses on multi-modal information retrieval, emphasizing feature fusion and cross-modal hashing models for the intelligent organization and retrieval of heterogeneous information resources.

Wenjie Meng received her Bachelor of Science in Communication Engineering and a Master's of Science degree in Computer Application Technology from the China University of Petroleum (East China), China, in 2004 and 2007, respectively. Her research interests include cross-modal information retrieval, with a focus on feature-based representation learning and hashing methods, with applications in multimedia information processing and intelligent resource construction.

Xuesong Zhang received his Bachelor's of Science degree in Library Science from Northeast Normal University, China, in 2005, and his Master of Science degree in Computer Science from the China University of Petroleum (East China), China, in 2016. He is currently an Associate Researcher at the Library of the China University of Petroleum (East China), China. His research focuses on information retrieval and knowledge services, particularly multimodal resource organization and data-driven management in large-scale, heterogeneous information environments.