

A Library Retrieval Model Using Feature Fusion and Cross-Modal Retrieval Techniques

Zhichao Li¹ and Fang Wang²

¹ Librarian, Information Consultation Department, Dezhou University, Dezhou, 253023, China, E-mail: dzxylzc@163.com (corresponding author).

² Associate Research Librarian, Discipline Inspection Commission, Dezhou University, Dezhou, 253023, China, E-mail: wangfang0058@163.com

Project Management

Received December 25, 2025; revised February 10, 2026; August 7, 2026; accepted September 8, 2026
Available online September 20, 2026

Abstract: This study proposes a library retrieval model that uses feature fusion and cross-modal retrieval techniques to solve the weak processing of multi-modal resources and the low retrieval efficiency of current library retrieval methods. The model first uses deep learning algorithms to extract features from multi-modal library resources. It then fuses extracted image features and text features. After that, it builds a similarity matrix that measures the relation between images and texts. The model uses this matrix to perform a hash function search and completes efficient image and text retrieval in the library. Feature extraction experiments show that the model achieves the highest mean average precision in cross-modal retrieval scenarios involving textual queries and visual inputs, as well as visual queries and textual outputs, reaching 0.8953 and 0.9034, respectively. The model also obtains the highest accuracy under various noise conditions, with a mean accuracy of 95.08%. These findings demonstrate the proposed model's ability to enhance the efficiency of library resource retrieval and facilitate more effective resource management and utilization.

Keywords: Feature fusion; cross-modal retrieval techniques; library retrieval model; deep learning; hash function.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2025-358

1. Introduction

The library retrieval system is a tool that helps users obtain library resources quickly and accurately, and its core function is to organize and index resource metadata to support users in locating target resources (Zhu et al., 2023). Traditional library retrieval methods rely on keyword matching and often exhibit low efficiency and low accuracy, failing to meet user needs (Wang et al., 2023). With advances in intelligent technologies, an increasing number of scholars have applied intelligent algorithms to library retrieval systems to enhance retrieval performance and accuracy (Hou et al., 2022). These technologies often focus on a single modality and only process text information. They ignore the rich semantic features of other modalities in library resources, leading to incomplete and inaccurate retrieval results (Li et al., 2024). Exploring multimodal retrieval techniques in library retrieval systems is therefore important. Feature fusion extracts and fuses features from library images and text, and multi-modal retrieval uses fused features to measure cross-modal similarity and complete retrieval tasks. This process reduces the limits of single-modal retrieval (Sun et al., 2024). Considering the limitations of existing library retrieval approaches, this paper proposes a library retrieval model that integrates feature fusion and cross-modal retrieval techniques to provide more intelligent services and better satisfy user requirements. The model is innovative because it integrates feature extraction, feature fusion, and hash retrieval modules. This integration overcomes the limitations of traditional models and enables efficient, accurate retrieval of multimodal information. The study focuses on the closed library corpus, whose resources are professionally curated but limited in scale, in contrast to open retrieval in networked environments. Based on this, two hypotheses are proposed: Hypothesis 1) suggests that the accuracy of cross-modal retrieval in closed library corpora correlates with feature fusion strategies. Hypothesis 2) suggests that optimizing feature mapping can reduce retrieval errors in closed scenarios. This paper expects that the model will become a useful solution to current problems in library retrieval and supports further development toward more intelligent systems. The innovation of this study is mainly reflected in two aspects: first, it integrates deep feature extraction, feature fusion and efficient hash-based retrieval for the first time, forming a complete technical process suitable for the library scene. Secondly, it overcomes the limitations of traditional single-mode retrieval and enables deep integration and accurate matching of multimodal information, such as images and text.

2. Related Works

Library retrieval tools can help users efficiently and accurately acquire high-quality, professionally screened resources, saving time on information filtering and improving information quality. With advances in intelligent technologies, more researchers have adopted intelligent algorithms in retrieval systems to improve performance and accuracy. For instance, Zhang et al. (2024) developed a privacy-preserving speech retrieval approach based on deep hashing and fully homomorphic encryption, which achieved an average accuracy exceeding 93% while enhancing data security. Ueno et al. (2022) introduced a cell retrieval approach utilizing light-triggered gas-producing polymer microarrays to enable targeted retrieval of individual cells and specific cell groups. The method efficiently isolated adherent cells with high selectivity. Wu et al. (2023) developed a cross-modal text-video retrieval framework that integrates attention enhancement and multi-task optimization strategies, achieving improved retrieval performance across several benchmark datasets. Huang et al. (2025) proposed an RFID-based intelligent retrieval method for library book resources. Experimental results showed that this method could reduce intelligent retrieval errors for library book resources. Mukhopadhyay et al. (2023) proposed using geodesic search to enhance library information retrieval systems and address the problem of insufficient information utilization. The proposed approach achieved notable improvements in library retrieval efficiency.

Feature fusion is the integration of information from multiple sources, dimensions, or modalities into a unified feature representation using specific strategies. Li and Zhou (2025) proposed an improved lightweight multi-scale feature fusion algorithm and applied it to face recognition on mobile devices. Experimental results showed that the feature fusion module could more effectively learn and integrate facial features at various scales to achieve efficient face recognition. Li et al. (2022) proposed an asymmetric feature integration network for classifying hyperspectral and synthetic aperture radar images. This network designed a feature calibration module that leveraged spatial dependencies among multi-source features to enhance the model's discriminative capability. Experimental results showed that this method outperformed comparison methods on three datasets. Chen and Mo (2023) proposed a deep learning-based recognition model incorporating attention mechanisms and feature fusion to enhance human action recognition performance. The proposed model achieved accurate recognition results. Cross-modal retrieval technology is an information retrieval technique that breaks down barriers between data modalities, enabling users to use one data modality as a query condition to retrieve other, semantically related modalities, playing an important role in multiple fields. Hu et al. (2023) introduced a cross-modal retrieval approach for handling partial data inconsistency. Comparative experiments across nine advanced methods verified their effectiveness in improving data-matching performance. Huo (2023) developed a label-guided cross-modal hashing framework that projected heterogeneous modal data into a shared semantic space to preserve meaningful information and reduce modality discrepancies. The method achieved higher retrieval accuracy and efficiency compared with existing approaches.

Overall, previous studies have achieved notable advances in retrieval technologies; however, challenges remain regarding retrieval efficiency and data processing capabilities. Feature fusion and cross-modal retrieval techniques offer potential solutions by enabling the selection of appropriate intelligent algorithms tailored to specific requirements. Therefore, the library retrieval model developed in this study, which integrates feature fusion and cross-modal retrieval techniques, aims to enhance retrieval efficiency and accuracy while achieving more precise semantic matching.

3. Library Retrieval Model Construction

3.1. Feature Extraction Algorithm Design Based on Deep Learning

The library mainly contains image resources and text resources. These resources are heterogeneous in format, and traditional single-modal retrieval methods only retrieve texts. Such methods cannot meet the need for multi-modal resources in the library (Marolt et al., 2022). Cross-modal retrieval techniques bridge the semantic gap between modalities and enable cross-modal matching. They support accurate joint retrieval of images and texts (Kale et al., 2022). Direct cross-modal retrieval over large numbers of images and texts increases computational cost and reduces retrieval efficiency (Tran et al., 2023). The model therefore extracts features from text and images before performing cross-modal retrieval. This step reduces computation in a clear way. The Bidirectional Gated Recurrent Unit (BiGRU) model captures contextual information in both directions and mitigates long-sequence forgetting. It also keeps a balance between efficiency and expression. The model, therefore, uses BiGRU to extract textual features from library texts. Fig. 1 shows the basic process.

As illustrated in Fig. 1, the algorithm processes library text inputs by segmenting sentences into individual words. A learnable embedding matrix is then applied to each word, transforming it into the embedding space and generating its corresponding vector representation. This step converts discrete words into continuous vectors, enabling subsequent neural network processing. The algorithm then passes the word embeddings through the BiGRU layer sequentially, completing bidirectional feature extraction. The calculation process is shown in Eq. (1).

$$\begin{cases} \vec{h}_j = GRU(e_j, \vec{h}_{j-1}) \\ \overleftarrow{h}_j = GRU(e_j, \overleftarrow{h}_{j-1}) \end{cases} \quad (1)$$

In Eq. (1), $\vec{h}_j$ and $\overleftarrow{h}_j$ represent forward and backward hidden states in GRU time steps. After BiGRU processing, the model integrates and transforms features in the connection layer, yielding the word representation shown in Eq. (2).

$$t_j = \frac{\vec{h}_j + \overleftarrow{h}_j}{2} \quad (2)$$

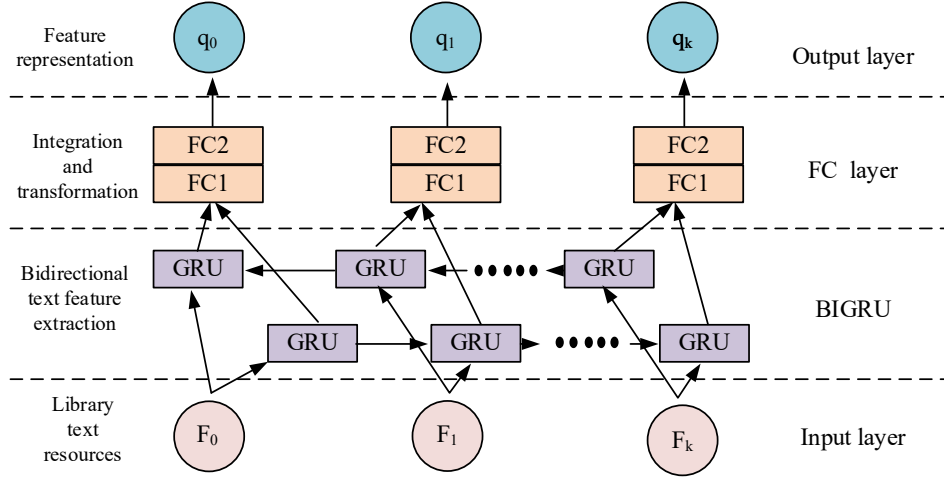


Fig. 1. BiGRU text feature extraction algorithm

Based on the word representation, the model then computes the global text representation. The process is shown in Eq. (3).

$$t^{(g)} = \frac{1}{n} \sum_{j=1}^n t_j \quad (3)$$

After extracting text features, the model continues to extract image features for better multi-modal integration. A Convolutional Neural Network (CNN) is employed to extract visual features from library images, and the activation function is selected to improve model performance. The process is shown in Eq. (4).

$$\begin{cases} f(x) = x & x > 0 \\ f(x) = ae^x - a & x \leq 0 \end{cases} \quad (4)$$

In Eq. (4), a is an adjustable parameter. The part greater than 0 solves the gradient vanishing problem, while the part less than 0 makes the activation function more robust to input changes, thereby improving training speed and efficiency. This improves training speed and efficiency. Based on this, the Improve CNN (ICNN) image feature extraction process is shown in Fig. 2.

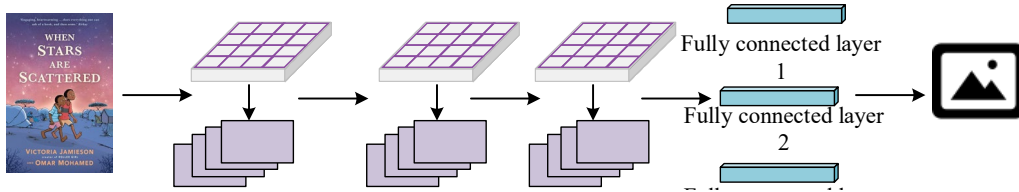


Fig. 2. ICNN image feature extraction process (Image source from: pixabay.com)

As shown in Fig. 2, the input library image is processed through convolution and pooling operations to obtain hierarchical visual representations. The image then enters the third convolution layer for deeper feature extraction, and then the third pooling layer to simplify the representation. The model then passes the data through three fully connected layers, producing the final image feature representation.

3.2. Feature Fusion Algorithm Design with Feature Extraction

In the library cross-modal retrieval task, image features and text features are usually trained with independent single-modal algorithms. They differ in feature dimension, distribution, and semantic representation (Pal et al., 2023). After the model extracts image and text features using BiGRU and ICNN, it designs a feature fusion module to fuse them. This step supports retrieval in a unified space. The Gated Feature-Residual Feature (GF-RF) method converts text features and image features and builds a unified structural representation. The model, therefore, uses GF-RF for multi-modal fusion. Fig. 3 shows the GF-RF module.

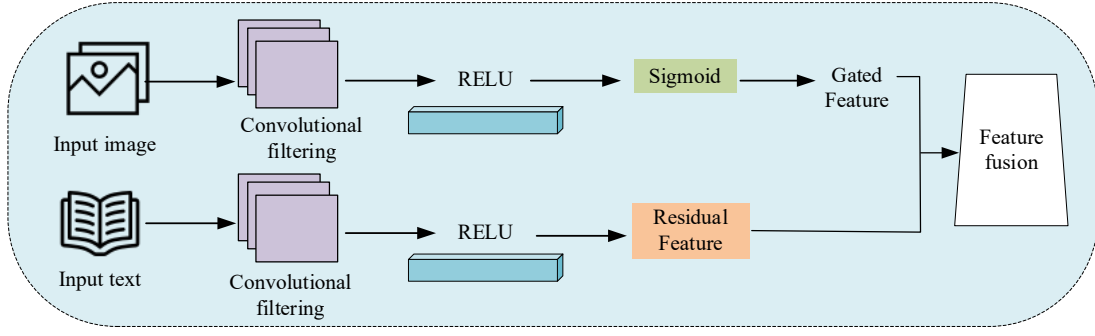


Fig. 3. GF-RF feature fusion module

In Fig. 3, the method processes both image and text features simultaneously. The model first builds gated features from image features. The process is shown in Eq. (5).

$$f_{gate}(\phi_x, \phi_t) = \sigma(W_g * ReLU(\phi_t)) \odot \phi_x \quad (5)$$

In Eq. (5), σ represents the sigmoid function, and W_g represents the convolution filter. The model then converts text features into residual features. The process is shown in Eq. (6).

$$f_{res}(\phi_x, \phi_t) = W_r * ReLU(\phi_t) \quad (6)$$

In Eq. (6), W_r is the convolution filter. After building gated and residual features, the model combines them linearly to balance the effects of single-modal features. The model then gets a better feature representation. The process is shown in Eq. (7).

$$\phi_{xt}^{rg} = W_g f_{gate}(\phi_x, \phi_t) + W_r f_{res}(\phi_x, \phi_t) \quad (7)$$

In Eq. (7), W_g and W_r represent weight information. After this combination, the model completes the fusion process. To measure the gap between the prediction and the target and to optimize parameters for accurate learning, the model uses the softmax cross-entropy loss function. The definition is shown in Eq. (8).

$$L = \frac{-1}{MB} \sum_{i=1}^B \sum_{m=1}^M \log \left\{ \frac{\exp\{k(\psi_i, \phi^+)\}}{\sum_{\phi_j \in N_i^m} \exp\{k(\psi_i, \phi^+)\}} \right\} \quad (8)$$

In Eq. (8), k represents a similarity kernel function, defined as the negative Euclidean distance. This function measures the difference between the prediction and the target. The model then combines the BiGRU text feature extraction algorithm, the ICNN image feature extraction algorithm, and the GF-RF fusion module and builds the final BiGRU-ICNN-GF-RF algorithm. Fig. 4 shows the process.

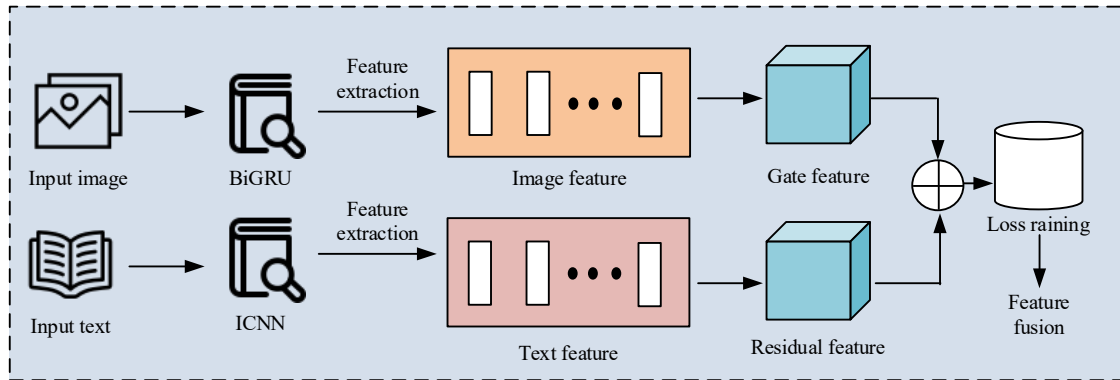


Fig. 4. BiGRU-ICNN-GF-RF feature fusion algorithm process

As shown in Fig. 4, the algorithm first receives image data and text data as inputs. It uses ICNN and BiGRU to extract image features and text features. It then uses the GF-RF module to convert image features into gated features and convert text features into residual features. The model combines these features to achieve multi-modal fusion. The fused features then enter the loss function for training and parameter optimization. This process supports robust and efficient multi-modal feature fusion.

3.3. Cross-Modal Retrieval Model Construction Based on Feature Fusion

The feature fusion algorithm integrates features from different modalities, providing a solid foundation for the library's cross-modal retrieval model. In real-world library applications, users expect a simple, direct way to access related multimodal resources (Wang et al., 2024). To meet this need, the model builds a cross-modal retrieval module using a

fusion algorithm. The Hash function maps high-dimensional features into low-dimensional binary Hash codes. The model then matches large-scale data with Hamming distance in a fast and efficient way. Hash-based methods have therefore become a common choice for cross-modal retrieval in recent years. Using feature fusion, the model performs cross-modal retrieval via a Hash learning module. Fig. 5 shows the Hash retrieval process in the library.

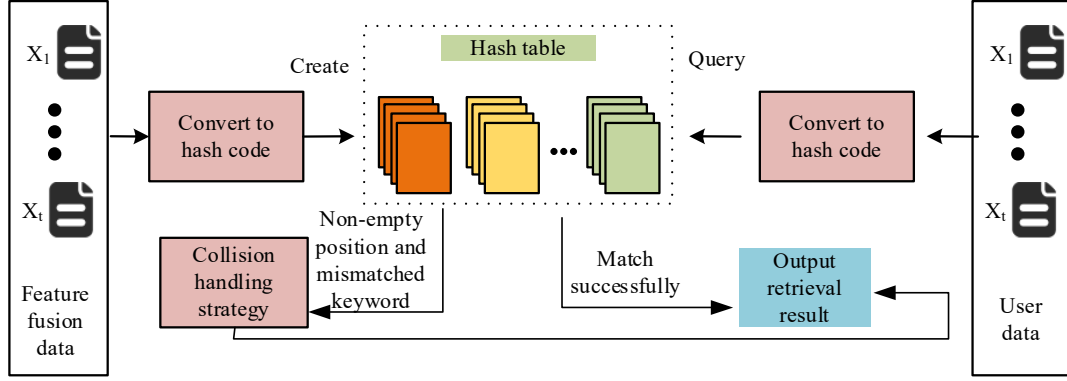


Fig. 5. Cross-modal Hash retrieval process for library resources

As shown in Fig. 5, the model first fuses features from different modalities and converts them into unified Hash codes. It then builds a Hash table to store target items. When a user sends a cross-modal query, the model extracts multi-modal features from the query and fuses them into a unified high-dimensional vector. It maps this vector into a Hash code through the same Hash function. It then locates the corresponding position in the Hash table. If the table contains a Hash code that matches the query, the model returns the related library item. If the position is empty or the code does not match, the retrieval fails. If the position is not empty but does not match the query, the model continues checking the next position according to the conflict-handling strategy until it finds a matching item. To support this process, the model includes a similarity-matrix module. This module receives fusion results and helps the Hash encoder increase similarity between multi-modal Hash codes. The model, therefore, builds a similarity matrix before Hash retrieval. The matrix includes cross-modal cosine similarity, cross-modal domain similarity, and intra-modal similarity. The model first computes cross-modal cosine similarity. This similarity fuses structural information from different modalities. The process is shown in Eq. (9).

$$S_d(x_i, y_i) = (1 - a) \cos(x_i, y_i) + a \cos(x_i, y_i) \quad (9)$$

In Eq. (9), a represents a parameter that adjusts modal weight and controls the ratio between the two modalities. The model then computes cross-modal domain similarity. This similarity captures second-order distance information and addresses the limitation of relying solely on first-order distance. The process is shown in Eq. (10).

$$P(\sigma(x_i, y_j)) = [q \in Nk(i)] S_d(x_i, y_i) \quad (10)$$

In Eq. (10), $\sigma(x_i, y_j)$ represents the similarity predicate of x_i and y_j and defines the logic of similarity. $Nk(i)$ represents the k neighbor value of x_i . The model then computes intra-modal similarity. This similarity keeps semantic consistency within the same modality. The process is shown in Eq. (11).

$$S_{dd} = \left(\frac{x_i \cdot \tilde{x}_i}{\|x_i\| \times \|\tilde{x}_i\|} \right) + \left(\frac{y_i \cdot \tilde{y}_i}{\|y_i\| \times \|\tilde{y}_i\|} \right) \quad (11)$$

In Eq. (11), x_i and $\tilde{x}_i$ represent text feature vectors before and after fusion. y_i and $\tilde{y}_i$ represent image feature vectors before and after fusion. The model sums these similarity values and gets the final similarity, as shown in Eq. (12).

$$S(x_i, y_i) = (1 - \beta) S_d + \beta S_{nn} + a S_{ff} + S_{dd} + S_{fph} \quad (12)$$

This value reflects the semantic relations between cross-modal data more accurately. The final similarity value stays within the range [0,1]. The model then builds the similarity matrix based on this range. After integrating the BiGRU-ICNN-GF-RF fusion algorithm, the similarity matrix module, and the Hash retrieval module, the model gets the final BiGRU-ICNN-GF-RF-Hash retrieval model. The short name is the BIGH model. Fig. 6 shows the full process.

As shown in Fig. 6, the model first enters the feature extraction module. It uses ICNN to process image resources and extract image features. It uses BiGRU to process text resources and extract text features. The model then passes both types of features to the fusion module, completing feature fusion. It then builds a similarity matrix to measure the relation between images and texts. Finally, based on this similarity matrix, the model completes Hash lookup. The Hash learning module converts high-dimensional features into compact Hash codes and supports efficient retrieval and matching between image and text resources in the library. In summary, a cross-modal retrieval model adapted to closed library corpora was constructed through feature extraction, feature fusion, and deep hash retrieval. In order to clearly understand the definitions of various formula symbols involved in the above methods, standardize the use of symbols throughout the text, and provide specific explanations for all formula symbols, as shown in Table 1.

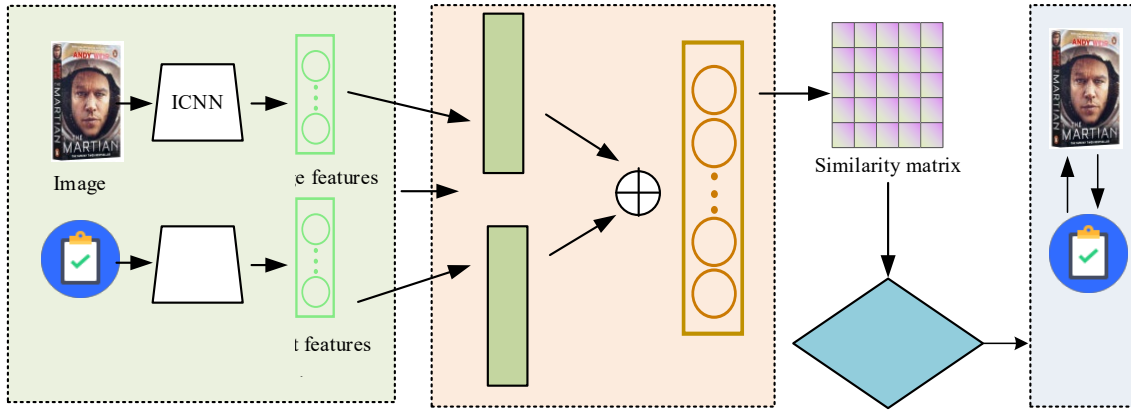


Fig. 6. BIGH model process (Icon source from: <https://pixabay.com/>)

Table 1. Explanation of formula symbols

Symbol	Meaning	Symbol	Meaning
$\vec{h}_f$	Forward GRU time step shadow hiding status	$\vec{h}_b$	Backward GRU time-step hidden state
n	Total number of words	a	Adjustable parameter
σ	Sigmoid function	W_g	Convolution filter
W_g	Weight information	W_r	Weight information
k	Similarity kernel function	$\odot$	Element-wise multiplication unit
$Nk(i)$	k -nearest neighbor of x_i	$\sigma(x_i, y_j)$	Similarity predicate of x_i and y_j
$\tilde{x}_l$	Fused text feature vector	x_i	Text feature vector before fusion
$\tilde{y}_l$	Fused image feature vector	y_i	Image feature vector before fusion

4. Performance Test of Feature Fusion Algorithm and Library Retrieval Model

To verify the performance of the BiGRU-ICNN-GF-RF feature fusion algorithm and the BIGH library cross-modal retrieval model proposed in this study, a systematic performance testing experiment was conducted. The detailed data collection and processing plan is as follows: Firstly, through the public crawling interfaces of the National Digital Library Museum and Douban Reading, targeted collection of classic book covers, academic journal illustrations, and other multi theme images is carried out, and the corresponding text descriptions such as book introductions and journal abstracts are synchronously obtained for each image; Secondly, strict filtering is performed on the raw data to remove samples with image resolution below 224×224 pixels and text description length less than 50 words. The semantic correlation between images and text is manually annotated and verified; only samples with correlation ≥ 0.8 are retained. The research ultimately constructed a dataset comprising 10,000 image-text pairs, which was divided into training, validation, and test sets in an 8:1:1 ratio. To ensure model reproducibility and transparency, the study details the architectural parameters of each module in the BIGH model. Specifically, the first convolutional layer of ICNN has 128 channels, a 3×3 kernel, a stride of 1, and an output size of $224 \times 224 \times 128$. The second convolutional layer has 256 channels, a 3×3 kernel, a stride of 1, and an output size of $112 \times 112 \times 256$. The BiGRU hidden layer has 256 dimensions and an output size of $512 \times 512 \times 128$. The GFRF module has 256 channels, a 1×1 kernel, a stride of 1, and an output size of $1 \times 1 \times 256$. The hash retrieval module has a hash code length of 128 and an output size of $1 \times 1 \times 128$.

4.1. Performance Test of BiGRU-ICNN-GF-RF Algorithm

To validate the effectiveness of the proposed BiGRU-ICNN-GF-RF feature fusion algorithm, the study compared it with three other algorithms, which include Term Frequency-Inverse Document Frequency-Residual Neural Network-50-Weighted Concatenation Fusion (TF-IDF-ResNet50-WCF), Bidirectional Encoder Representations from Transformers-Long Short-Term Memory-Attention Mechanism Fusion (BERT-LSTM-AMF), and Bidirectional Encoder Representations from Transformers-Back Propagation Neural Network-Cross-Modal Interaction Fusion (BERT-BP-CMIF). First, the study compared the accuracy and F1 scores of the feature extraction. The comparison results are shown in Fig. 7.

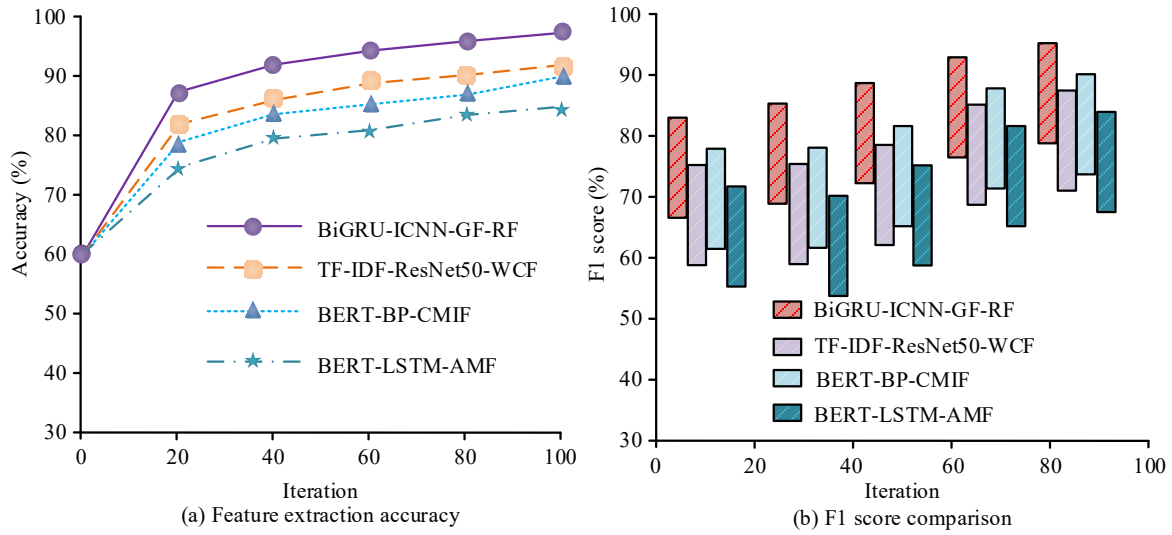


Fig. 7. Feature extraction accuracy and F1 scores of the four algorithms

Feature extraction accuracy is a metric that evaluates a model's ability to extract meaningful features from raw data. The more effective the extracted features, the higher the probability of correct matching in downstream tasks. As shown in Fig. 7(a), the BiGRU-ICNN-GF-RF algorithm consistently achieved the highest feature-extraction accuracy, reaching 97.8% upon completion of the iterations, significantly higher than the comparison algorithms' 91.5%, 90.2%, and 85.7%. As shown in Fig. 7(b), the F1 scores of BiGRU-ICNN-GF-RF were all above 80% and reached 95.1% upon completion of the iterations, significantly higher than those of the comparison algorithms. In summary, the proposed algorithm demonstrated excellent performance in extracting features from library text and image resources, establishing a solid and reliable foundation for subsequent feature fusion and resource retrieval matching. Next, the study compared the visualization results after feature extraction by the four algorithms. The comparison results are shown in Fig. 8.

In Fig. 8, different colors represented different feature types. The more concentrated the colors and the clearer the boundaries, the better the feature extraction effect. In Fig. 8(a), after the BiGRU-ICNN-GF-RF algorithm extracted features, the feature distribution points were relatively concentrated, and the distinction among categories in the two-dimensional space was clearer. As shown in Figs. 8(b), (c), and (d), after feature extraction by the other three algorithms, the feature distribution points were relatively scattered, and the distinction in the two-dimensional space was not sufficiently clear. These results demonstrated that the features extracted by the BiGRU-ICNN-GF-RF algorithm exhibited greater discriminability and maintained large inter-class differences after dimensionality reduction, which was more beneficial for subsequent feature-based analysis or task execution. Next, the study compared the loss curves and average similarity of feature fusion. The comparison results are shown in Fig. 9.

As shown in Fig. 9(a), the loss of the BiGRU-ICNN-GF-RF algorithm decreased rapidly with increasing training epochs, dropping to 0.28 at 10 epochs and then continuing to decrease, ultimately stabilizing at 0.05. In contrast, the loss values of the other three algorithms decreased more slowly and remained above 0.3 at 20 epochs. As shown in Fig. 9(b), the average similarity after feature fusion of the BiGRU-ICNN-GF-RF algorithm reached 95.4% after iteration completion, significantly higher than the other comparison algorithms. These results indicated that the BiGRU-ICNN-GF-RF algorithm not only achieved good convergence performance but also fully explored the correlation between text and images, making the fused features better reflect the overall attributes of books.

4.2. Performance Verification of Library Retrieval Model

Following validation of the feature fusion algorithm, the study continued performance testing on the BIGH library cross-modal retrieval model. The study compared it with three other cross-modal retrieval models: Graph Convolutional Neural-Attention Mechanism (GCN-AM), Canonical Correlation Analysis-Adversarial Learning (CCA-AL), and Residual Neural Network-Gated Recurrent Unit (ResNet-GRU). First, the study compared the Precision-Recall (PR) curves of the four models. The comparison results are shown in Fig. 10.

In Fig. 10(a), in the text-to-image retrieval task, the BIGH model achieved the largest area under the curve (AUC), with an Average Precision (AP) of 0.8953, while the AP values of the other models were all below 0.86. In Fig. 10(b), in the image-to-text retrieval task, the BIGH model maintained its lead in both recall and precision, achieving an AP of 0.9034, significantly higher than the comparison models. In summary, in both text-to-image and image-to-text retrieval tasks, the BIGH model achieved the highest AP values and optimal performance. Next, the study compared the retrieval time and memory usage of the four models. The comparison results are shown in Fig. 11.

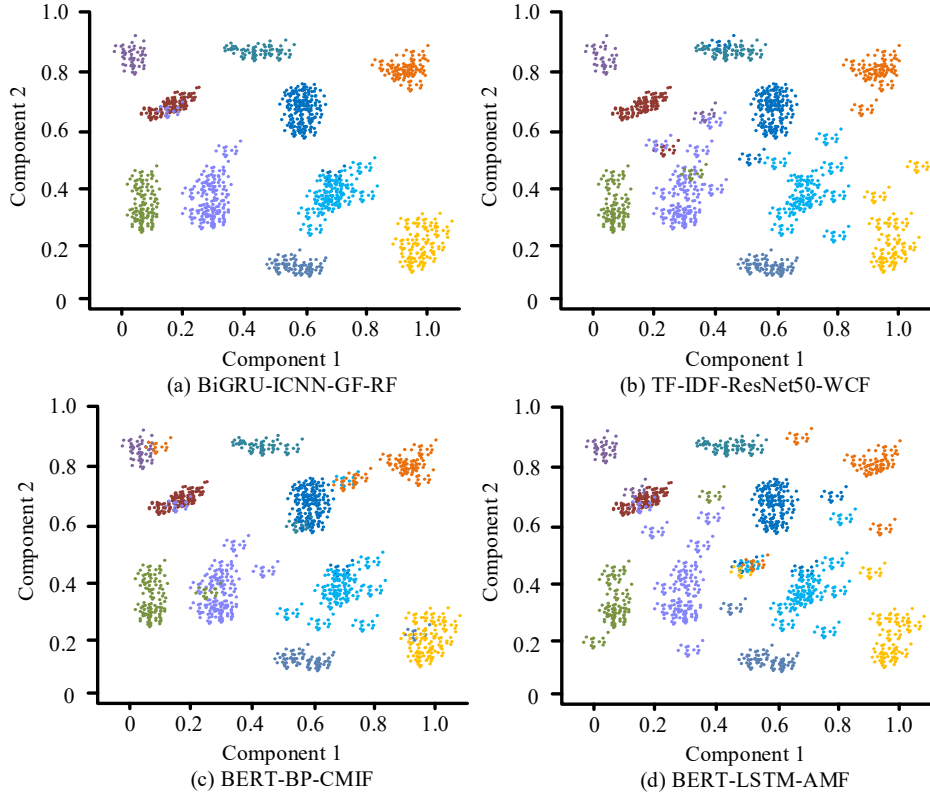


Fig. 8. Visualization results after feature extraction by the four algorithms

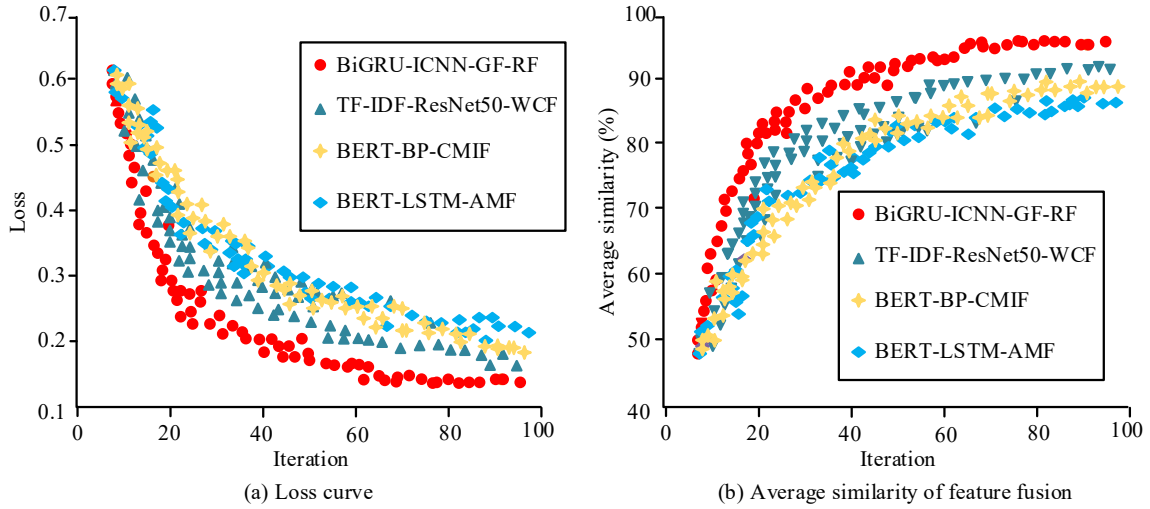


Fig. 9. Loss curves and average similarity of feature fusion

As shown in Fig. 11(a), when the data volume increased to 3,000, the BIGRH model's retrieval time was only 17.3s, outperforming the other models in retrieval efficiency. As shown in Fig. 11(b), the memory space occupied by the BIGRH model increased slowly and ultimately stabilized at around 290.4 MB, significantly lower than that of the comparison models. Overall, the BIGRH model showed minimal performance fluctuations when processing large volumes of financial data and demonstrated more stable, efficient data processing capabilities compared to the comparison models. Finally, the study compared the retrieval accuracy of the four models under different noise conditions. The retrieval accuracy of the four models under different noise conditions is compared, and the results are shown in Table 2.

In Table 2, the BIGRH model achieved the highest retrieval accuracy across all noise conditions, with an average accuracy of 95.08%, significantly higher than the other three models (91.26%, 89.02%, and 87.66%). These results demonstrated that the BIGRH model was robust in cross-modal retrieval and provided an efficient, stable solution for the library retrieval field. To validate the functions of each module in the BIGRH model, the study conducted an ablation experiment, with results presented in Table 3. According to Table 3, when the ICNN module, GF-RF fusion layer, and Hash layer are sequentially removed from the model, the model's accuracy, recall, F1 score, and precision all decrease

significantly. After removing the ICNN module, the accuracy decreased by 2.7%, the recall decreased by 2.9%, the F1 value decreased by 2.8%, and the accuracy decreased by 2.6%. After removing the GF-RF fusion layer, the decreases across various indicators widened further, with a 5.1% decrease in accuracy compared to the original model, a 5.3% decrease in recall, a 5.2% decrease in F1, and a 4.9% decrease in F1. After removing the hash layer, the performance degradation was most significant, with accuracy 85.6%, recall 84.8%, F1 85.2%, and accuracy 87.8%. This indicated that the ICNN module, GF-RF fusion module, and hash module all played important roles in improving cross-modal retrieval performance. To conduct a comprehensive evaluation of the model's performance, the study compared the BIGRH model against three cross-modal baselines: Contrastive Language-Image Pre-training (CLIP), Aligning Language and Vision with BERT (ALBEF), and Foundational Language and Vision Alignment (FLAVA). The comparison results are presented in Table 4.

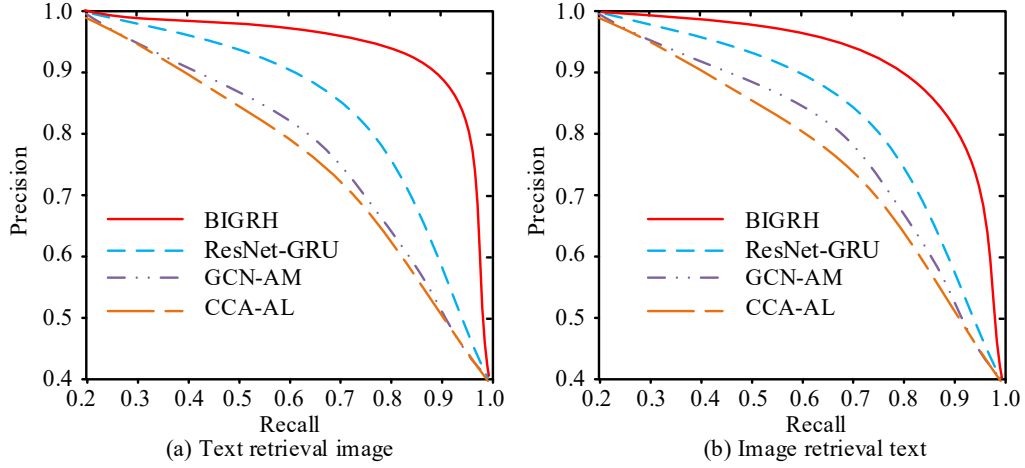


Fig. 10. PR curves for text-to-image retrieval and image-to-text retrieval

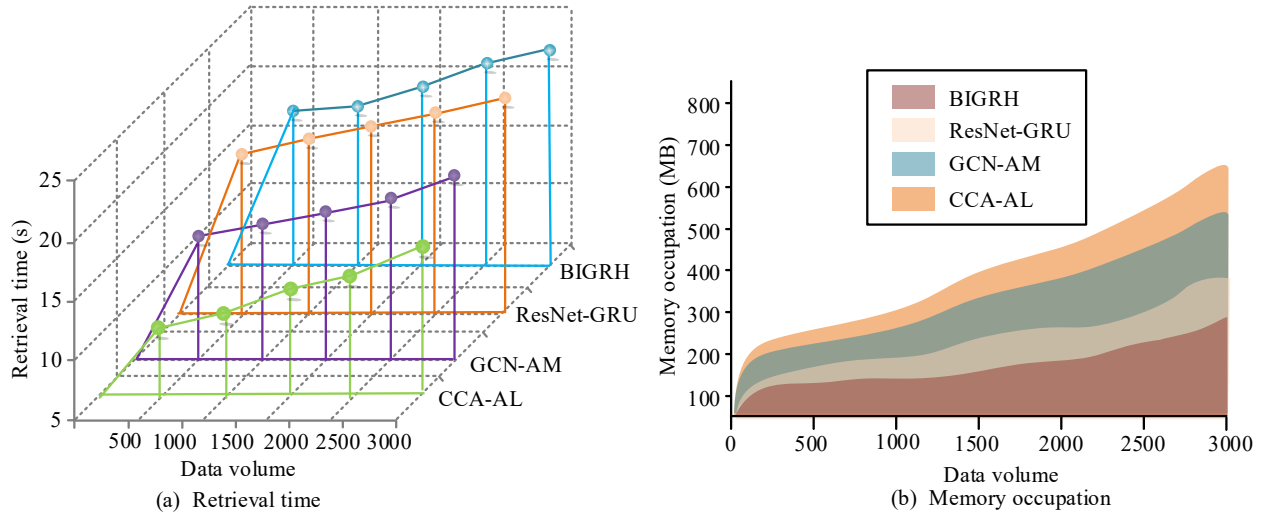


Fig. 11. Retrieval time and memory usage of the four models

Table 2. Retrieval accuracy of the four models under different noise conditions (%)

Noise type/Model type	BIGRH	ResNet-GRU	GCN-AM	CCA-AL
Blurred image	94.29	90.41	88.23	87.69
Incomplete image	94.18	90.26	89.27	87.63
Illumination-distorted image	94.58	91.29	89.36	86.36
Erroneous text	95.38	91.13	88.36	86.29
Malformatted text	97.47	93.58	90.19	90.68
Text with disturbing vocabulary	94.57	90.88	88.72	87.30

Table 3. Ablation experiment results

Model	Precision (%)	Recall (%)	F1 value (%)	Accuracy
BIGRH	92.4	91.8	92.1	94.16
BiGRU-ResNet-GF-RF-Hash	89.7	88.9	89.3	91.5
BiGRU-ICNN-Concat-Hash	87.3	86.5	86.9	89.2
BiGRU-ICNN-GF-RF-FC	85.6	84.8	85.2	87.8

Table 4. Comparison results of BIGRH with cross-modal baseline model

Model name	Image retrieval R@1 (%)	Image retrieval R@5 (%)	Text retrieval R@1 (%)	Text search R@5 (%)
CLIP	78.6	91.2	74.3	88.5
ALBEF	82.4	93.7	79.1	90.6
FLAVA	79.8	92.5	76.5	89.3
BIGRH	85.2	95.7	84.6	96.4

According to Table 4, the BIGRH model achieved an R@1 index of 85.2% in image retrieval, which is 6.6%, 2.8%, and 5.4% higher than the three comparison models, respectively. The R@5 index for image retrieval is 95.7%, which is 4.5%, 2.0%, and 3.2% higher than CLIP, ALBEF, and FLAVA, respectively. In the text retrieval task, the BIGRH model achieved an R@1 of 84.6%, while the three comparison models were all below 80%. Text retrieval R@5 has reached 96.4%, 7.9% higher than CLIP, 5.8% higher than ALBEF, and 7.1% higher than FLAVA. This fully demonstrated that the BIGRH model outperformed current mainstream baseline models on cross-modal retrieval tasks, whether for image retrieval of text or text retrieval of images.

5. Conclusion

To address the problem that current library retrieval models largely rely on single-modal, low-efficiency methods, the study proposed a model that leverages feature fusion and cross-modal retrieval. The model first extracted text and image features from library data using BiGRU and ICNN, then used the GF-RF module to deeply fuse the two features, and finally applied a Hash function for retrieval. After completing the iterations, the algorithm achieved a feature extraction accuracy of 97.8% and an F1 score of 95.1%. Meanwhile, the loss of value decreased rapidly with increasing training epochs and eventually stabilized at 0.05. From a practical application perspective, the proposed model achieved the highest AP values and maintained superior retrieval accuracy across various noise conditions, reaching 97.47% in long-text scenarios. In conclusion, the cross-modal fusion retrieval model developed in this study overcomes the limitations of traditional single-modal technologies, offering a novel approach to intelligent library resource retrieval and advancing library services toward smarter, more precise operations. Although the proposed model demonstrated excellent performance in practice, its retrieval speed still requires further optimization for large-scale data. Therefore, future research can explore more efficient retrieval strategies to better meet libraries' diverse retrieval needs. Research has shown that cross-modal feature fusion and hash retrieval can significantly improve the accuracy and robustness of library retrieval. Managers can optimize decision-making from three aspects: integrating image feature extraction and hash indexing into resource collection and system upgrades to complement traditional text-based indexing; incorporating AP values and response time of cross-modal retrieval into service evaluation while providing image-text retrieval interfaces; and establishing a semantic consistency verification mechanism to ensure the relevance and quality of incoming resources. These improvements can facilitate the development of library retrieval systems toward greater intelligence and accuracy.

Author Contributions

Zhichao Li contributed to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Fang Wang contributed to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

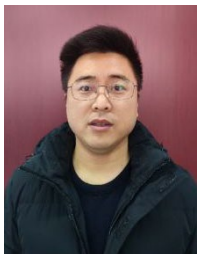
Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors confirm that no AI tools were used in the preparation of this manuscript.

References

- Chen, T. and Mo, L. (2023). Swin-fusion: Swin-transformer with feature fusion for human action recognition. *Neural Processing Letters*, 55(8), 11109–11130. doi:10.1007/s11063-023-11367-1
- Hou, C., Li, Z., and Wu, J. (2022). Unsupervised hash retrieval based on multiple similarity matrices and text self-attention mechanism. *Applied Intelligence*, 52(7), 7670–7685. doi:10.1007/s10489-021-02804-6
- Huang, W. and Ling, J. (2025). Intelligent retrieval method of book resources in smart library based on RFID. *International Journal of Reasoning-based Intelligent Systems*, 17(1), 32–39.
- Hu, P., Huang, Z., Peng, D., Wang, X., and Peng, 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
- 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. doi:10.1109/TCSVT.2023.3285266
- Kale, M., Dash, J., and Mukhopadhyay, S. (2022). Efficient image retrieval system for textural images using fuzzy class membership. *Multimedia Tools and Applications*, 81(26), 37263–37297. doi:10.1007/s11042-022-13529-0
- 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. and Zhou, Y. (2025). Bideeplab: An improved lightweight multi-scale feature fusion Deeplab algorithm for facial recognition on mobile devices. *Computer Simulation in Application*, 3(1), 57–65. doi:10.18063/csa.v3i1.917
- Li, W., Gao, Y., Zhang, M., Tao, R., and Du, Q. (2022). Asymmetric feature fusion network for hyperspectral and SAR image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10), 8057–8070. doi:10.1109/TNNLS.2022.3149394
- Marolt, J., Kosanić, N., and Lerher, T. (2022). Relocation and storage assignment strategy evaluation in a multiple-deep tier captive automated vehicle storage and retrieval system with undetermined retrieval sequence. *International Journal of Advanced Manufacturing Technology*, 118(9), 3403–3420. doi:10.1007/s00170-021-08169-x
- Mukhopadhyay, M. and Mukhopadhyay, P. (2023). From textual search to geodetic search: Enhancing library retrieval systems. *Indian Journal of Information Library & Society*, 36(2), 10–12. doi:10.5281/ZENODO.8222118
- Pal, S., Roy, A., Shivakumara, P., and Pal, U. (2023). Adapting a Swin Transformer for license plate number and text detection in drone images. *Artificial Intelligence and Applications*, 1(3), 145–154. doi:10.47852/bonviewAIA3202549
- Sun, Y., Dai, J., Ren, Z., Li, Q., and Peng, D. (2024). Relaxed energy preserving hashing for image retrieval. *IEEE Transactions on Intelligent Transportation Systems*, 25(7), 7388–7400. doi:10.1109/TITS.2024.3351841
- Tran, N. H., Nguyen, K. N., and Vo, V. H. (2023). Image retrieval based on ConvNets and hashing algorithm. *VNUHCM Journal of Science and Technology Development*, 26(4), 3048–3059. doi:10.32508/stdj.v26i4.4193
- Ueno, H., Akagi, Y., and Yamamura, S. (2022). Selective cell retrieval method using light-responsive gas-generating polymer-based microarrays. *Lab on a Chip*, 22(8), 1498–1507. doi:10.1039/d1lc01165k
- Wang, T., Zhu, L., Zhang, Z., Zhang, H., and Han, J. (2023). Targeted adversarial attack against deep cross-modal hashing retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10), 6159–6172. doi:10.1109/TCSVT.2023.3263054
- Wang, Z., Gao, Z., Yang, Y., Wang, G., Jiao, C., and Shen, H. T. (2024). Geometric matching for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3), 5509–5521. doi:10.1109/TNNLS.2024.3381347
- Wu, X., Qian, J., and Wang, T. (2023). Text-video retrieval method based on enhanced self-attention and multi-task learning. *Multimedia Tools and Applications*, 82(16), 24387–24406. doi:10.1007/s11042-023-14589
- Zhang, Q., Wen, Y., Huang, Y., and Li, F. (2024). Secure speech retrieval method using deep hashing and CKKS fully homomorphic encryption. *Multimedia Tools and Applications*, 83(26), 67469–67500. doi:10.1007/s11042-024-18113-2
- Zhu, L., Zheng, C., Guan, W., and Yang, Y. (2023). Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1), 239–260. doi:10.1109/TKDE.2023.3282921



Zhichao Li obtained his Master of Management degree from Shandong University of Technology in 2012. Currently, he works at the Library of Dezhou University as Director of the Information Consultation Department, holding the professional title of Librarian. He has published nine academic papers and authored one monograph to date. His main research interests focus on information resource management and library and information analysis.



Fang Wang is currently employed at the Discipline Inspection Commission of Dezhou University, with the professional title of Associate Research Librarian. She has authored two monographs and published several academic papers, and her current main research interest is Electronic Archives Management.