https://doi.org/10.31449/inf.v49i36.12357 Informatica **49** (2025) 171–184 **171**

A Deep Multimodal Retrieval Framework for Digital Libraries Using SE-ResNet-FCN, BERT, and Enhanced Hash Learning

Cheng Qiu
Zhejiang Library, Hangzhou 310008, China
E-mail: qiucheng7606@163.com

Keywords: fully convolutional network, hash learning, bidirectional transformer encoder representation, multimodal retrieval, attention mechanism

Received: October 10, 2025

The current multimodal retrieval accuracy of online libraries is insufficient. To solve this problem, a multimodal retrieval model for digital libraries that integrates fully convolutional networks and hash learning is proposed in the proposed method. The research introduces a fully convolutional network and a bidirectional Transformer encoder to extract semantic features, and combines a residual neural network to deeply optimize the model, thereby enhancing the feature expression ability. In the hash learning stage, a triplet loss and contrastive learning loss optimization model is designed to further enhance cross-modal semantic alignment. This enables image-text multimodal retrieval. In the experiment, the model used was applied to the BookCoverDataset for verification. And it is combined with Latent semantic sparse hashing (LSSH) and Collective Matrix Factorization Hashing. CMFH, Supervised Matrix Factorization Hashing (SMFH), Discrete online cross-modal hashing the multimodal retrieval models of digital libraries constructed by DOCH were compared. The experimental results show that the average retrieval accuracy score of this model is up to 0.93, and the maximum mAP reaches 0.95, which is 0.20 higher than that of the comparison model. When the hash code is 256 bits, its average accuracy reaches 0.94. Compared with the baseline model, the study proposes that the model demonstrates stronger semantic association ability and feature compression efficiency in multimodal retrieval tasks, verifying the effectiveness of the fusion strategy of fully convolutional networks and hash learning. The model significantly enhances the accuracy and robustness of cross-modal retrieval through a deep semantic alignment mechanism, providing a feasible solution for efficient and precise image-text mutual inspection in the digital library environment.

Povzetek: Študija predstavi model za boljše iskanje med slikami in besedilom v digitalnih knjižnicah, ki z združevanjem globokega učenja in "hash" kodiranja izboljša natančnost ter prekaša primerjalne metode.

1 Introduction

In the wave of digitalization, online libraries become the core platform for knowledge dissemination and sharing, with book resources presenting multimodal forms such as images and text [1]. Multimodal retrieval links heterogeneous data to meet users' precise demands of "searching text by image and searching image by text" and is crucial for improving digital resource utilization [2]. However, traditional retrieval methods focus on single-modal features and ignore cross-modal semantic associations. When handling large-scale multimodal data, they face high feature dimensionality, low retrieval efficiency, and difficulty in semantic alignment, making them unsuitable for the complex needs of online libraries [3]. Currently, multimodal retrieval research has made certain progress. Scholars attempt to link multimodal data through feature concatenation and early fusion, but they do not fully explore deep semantics between modalities. Existing models for online library multimodal resources often suffer from large semantic gaps and long retrieval time, failing to balance efficiency and accuracy [4-5]. The existing methods still have significant deficiencies in the collaborative optimization of cross-modal semantic alignment and feature compression, especially lacking the joint modeling ability of fine-grained semantic matching and efficient hash coding in the online library scenario. Moreover, most models fail to take into account the local feature alignment and global semantic consistency of image and text modalities, resulting in limited accuracy of retrieval results. To this end, a multimodal hash framework based on a fully convolutional network and a bidirectional Transformer encoder is proposed. A cross-modal attention mechanism is introduced to achieve fine-grained alignment of image regions and text primitives. The hash encoding process is optimized by combining semantic preservation loss and quantization constraints, significantly improving computational efficiency while ensuring retrieval accuracy.

The significance of the research lies in its supporting role in the efficient organization and precise access of multimodal resources in online libraries, which is directly related to the efficiency of users' information acquisition and service quality. The criteria for measuring success include a significantly better cross-modal retrieval accuracy than existing methods and an increase of no less than 5% in mAP values on mainstream evaluation

datasets. When the hash encoding compression ratio is below 64 bits, it still maintains a relatively high retrieval accuracy. The response time of the model meets the real-time requirements, and the time consumption of a single query is controlled within 200 milliseconds. It has good generalization ability and stability in the real online library scenario.

The research contribution lies in (1) proposing a cross-modal hash learning framework that integrates a fully convolutional network with a bidirectional Transformer encoder to achieve deep semantic alignment of image and text modalities at a fine-grained level; (2) Design a cross-modal attention fusion mechanism and a hybrid loss function, and collaboratively optimize the semantic preservation and hash quantization processes to enhance the distinguishability of multimodal features in a compact binary space; (3) Verify the model's effectiveness in both public datasets and real online library scenarios, balancing high-precision retrieval and low-latency response, and provide feasible solutions for the efficient organization of multimodal resources.

2 Literature review

Multimodal retrieval is a technique that uses a query from one modality to find semantically related data in other modalities, such as text, images, or videos. This technique has been widely applied in many fields, and many scholars have conducted related studies. Deep cross-modal hashing methods have been developed to learn fine-grained similarities between data points of different modalities, thereby improving retrieval accuracy. Contrastive learning frameworks, as an effective technique for enhancing the consistency of visual and text features, have been widely applied [6]. Knowledge distillation technology is also used to transfer knowledge from large teacher models to more effective student hash models, thereby reducing complexity while improving performance [7]. To address the challenge of efficiency, asymmetric hashing methods were proposed, with learning tasks defined within the framework of matrix factorization to generate compact and discriminative hash codes [8]. All these technologies have been applied in multimodal retrieval, and the research results have provided new improvement ideas for this field.

Methods such as FCN, hash learning, and image segmentation are also frequently employed in multimodal retrieval. With the advancement of science and technology, an increasing number of scholars are seeking more effective retrieval methods [9-10]. Zhang D et al. proposed a new online hashing method in order to effectively handle online streaming media data. This method utilizes semantic autoencoders to establish the correlation between binary codes and labels, and adopts the inner product of labels to achieve the connection between data and new data. This enhances the efficiency of large-scale cross-media similarity retrieval [11]. Khan A et al. proposed a cross-modal recovery technology based on a multi-label information depth ranking model to solve the problem of inappropriate information contained between images and texts in cross-modal retrieval. This technology uses a regularization function instead of binary constraints to confine discrete values within a numerical range for end-to-end training. The results show that the performance of this method on the MIR-Flickr-25K and NUS-WIDE datasets is significantly better than that of the existing mainstream models [12]. Wang Y et al. proposed a new supervised cross-modal hashing method, namely multi-information embedding hashing, in response to the problem that the retrieval performance of supervised cross-modal hashing is gradually reaching a bottleneck at present. Multi-information embedded hashing can flexibly handle various information mining, hash code learning and hash function learning, thus improving the retrieval performance of supervised cross-modal hashing [13]. Wen H et al. proposed a self-trained enhanced multi-factor matching network, which models the fine-grained alignment relationship between text and images by decoupling latent semantic factors and introducing a double aggregation mechanism. The results show that this method achieves significant performance improvements on multiple combined image retrieval benchmarks, with an improvement rate exceeding 8%[14]. To further compare the performance differences between the research method and the existing mainstream methods, the differences between each literature work and the research work, as well as their own advantages, are summarized in Table 1.

Table 1: Summary of the differences between the related work and the work of SRF-BERT-IDHS.

Document Number	Method	The difference from SRF-BERT-IDHS
[6]	A retrieval method combining deep neural model learning modalities and WOA operators	Deep semantic alignment has not been achieved. Only modal association is carried out in the shallow feature space, lacking cross-modal semantic consistency constraints.
[7]	A multimodal contrastive knowledge distillation method	The model has a high degree of redundancy and low computational efficiency
[8]	A novel asymmetric supervised fusion-oriented hashing method named ASFOG	Lack of dynamic adaptability
[11]	Cross-modal retrieval framework based on online hashing and semantic autoencoding	The semantic relationships between classes were not fully explored, resulting in limited generalization ability
[12]	Cross-modal recovery technology based on multi-label information deep ranking model	The semantic correlation is not high and the retrieval accuracy is low
[13]	A new supervised cross-modal hashing method	Insufficient modeling of fine-grained semantic associations between modalities
[14]	Self-trained enhanced multi-factor matching network	Relying on strongly supervised signals makes it difficult to address the challenges of label noise and modal heterogeneity

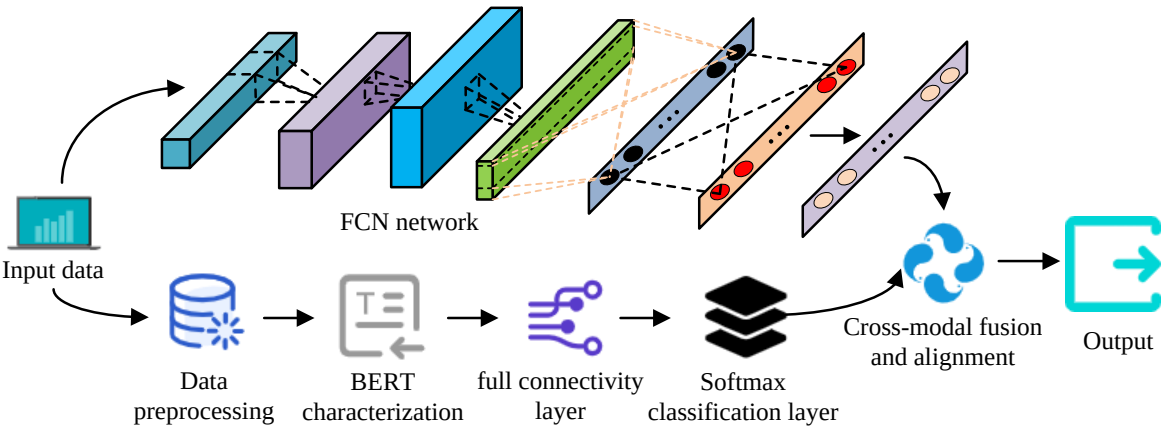


Figure 1: Modal data processing module composed of FCN and BERT.

Based on the above content, it can be known that although there has been certain progress in the related research of multimodal retrieval in digital libraries, due to the significant structural differences between the visual features of images and the language features of texts, deep semantic alignment has not been achieved, resulting in an imbalance between retrieval efficiency and accuracy. Moreover, most of the existing methods are limited to shallow semantic matching and it is difficult to capture the complex associations between modalities. Therefore, the proposed multimodal retrieval model for digital libraries, which integrates FCN and Hash Learning, replaces the original FCN network with Residual Network (ResNet) and combines a Squeeze-Excitation (SE) channel attention mechanism to enhance feature extraction. Using a dual-branch deep hashing network structure and an improved loss function, it aims to meet users' needs for accurate retrieval in digital libraries.

3 Construction of multimodal retrieval model for online libraries

3.1 Construction of modality feature extraction module based on FCN and BERT

Multimodal retrieval in online libraries aims to achieve precise content matching across or within modalities, such as "text→image," "image→text," and "image→image." To achieve this goal, it is necessary to address the significant differences in underlying representations among different modalities. In addition, features of different modalities need to be semantically aligned, and retrieval must balance efficiency and accuracy [15]. Therefore, SRF-BERT-IDHS proposes an online library multimodal retrieval model based on FCN and Hash Learning. The model uses FCN to process and analyze image modality data, employs Bidirectional Encoder Representations from Transformers (BERT) to capture textual features, and uses Hash Learning to address retrieval efficiency and large-scale data matching. The structure of the modality data processing module composed of FCN and BERT is shown in Figure 1.

As shown in Figure 1, after receiving multimodal raw data, images enter the FCN for encoding and decoding to achieve spatial feature alignment. Text enters BERT, which outputs context-aware word vectors through the Transformer encoder. Finally, the model performs cross-modal fusion and alignment of text and images, realizes cross-modal interaction, and outputs the results. In FCN, the convolution operation is the foundation for extracting image features. Its mathematical expression is shown in Equation (1) [16].

$$Y(i,j,c)=\sum_{p=0}^{k-1}\sum_{q=0}^{k-1}\sum_{c'=0}^{C_{in}-1}X(i+p,j+q,c')\cdot K(p,q,c',c)+b(c) \quad (1)$$

In Equation (1), $Y(i,j,c)$ denotes the pixel value of the output feature map. $X(i+p,j+q,c')$ is the pixel value of the input feature map at the corresponding position and channel. $K(p,q,c',c)$ is the weight of the convolution kernel at dimension (p,q,c',c). FCN reduces feature map resolution through pooling layers to enlarge the receptive field and reduce parameters. Its equation is shown in Equation (2).

$$Y(i,j,c)=\max_{p=0}^{s-1}\max_{q=0}^{s-1}X(i\cdot s+p,j\cdot s+q,c) \quad (2)$$

In Equation (2), s denotes the pooling window size, and $Y(i,j,c)$ is the maximum value. The core innovation of FCN is upsampling low-resolution feature maps to the input image size through deconvolution to achieve pixel-level prediction. The deconvolution operation is shown in Equation (3).

$$Y'(i,j,c)=\sum_{p=0}^{k'-1}\sum_{q=0}^{k'-1}\sum_{c'=0}^{C_{in}-1}X(i',j',c')\cdot K'(p,q,c,c')+b'(c) \quad (3)$$

In Equation (3), $i=i'\cdot s'-p+(k'-s')$, $j=j'\cdot s'-q+(k'-s')$, and $Y'\in R^{H\times W\times C_{out}}$ denote the high-resolution feature map after upsampling. The input representation in BERT is shown in Equation (4).

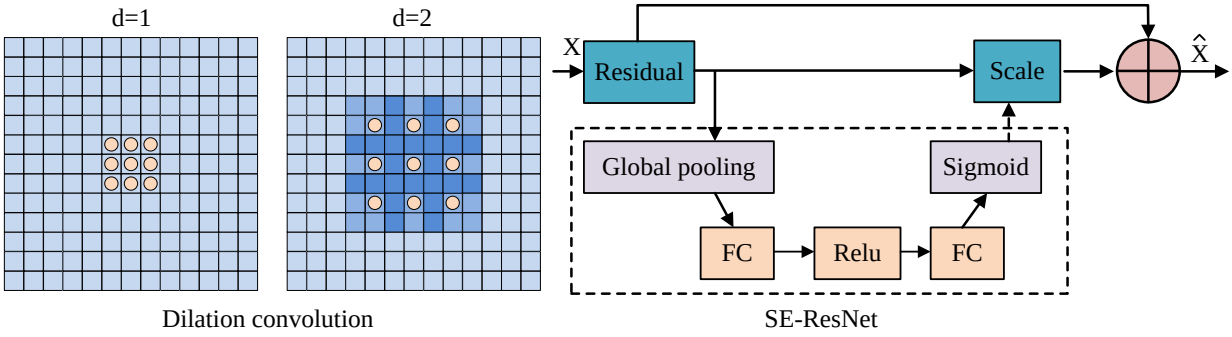


Figure 2: Diagram of dilated convolution and SE-ResNet structure.

$$E_i = T_i + P_i + S_i \tag{4}$$

In Equation (4), E_i denotes the final input vector of the i -th token. T_i is the word embedding, P_i is the position embedding, and S_i is the segment embedding. Using only FCN-BERT for online library multimodal retrieval can handle cross-modal information processing and matching to some extent. However, FCN has limited ability to capture small targets and fine details, lacks explicit modeling of global image semantics and spatial relationships, and consumes significant computational resources and time. Therefore, The proposed method replaces the VGG network in FCN with ResNet. The residual connections in ResNet solve the gradient vanishing problem in deep networks, support deeper network structures, and retain richer multi-scale features. Dilated convolution expands the receptive field without decreasing resolution, enabling the capture of the global layout of book covers. Attention mechanisms are applied by adding SE channel attention after each residual block and attention in BERT to focus on key information [17]. The schematic diagram of dilated convolution and SE-ResNet structure is shown in Figure 2.

As shown in Figure 2, the left diagram depicts dilated convolution, enlarging the receptive field while keeping the number of parameters unchanged. When d=1, the receptive field is 3×3. After dilation, it expands to 5×5, reducing overall computation. The right diagram shows the SE-ResNet structure, where the SE module is added to the residual structure, allowing the network to automatically focus on important feature channels and suppress irrelevant background. The output of dilated convolution is expressed in Equation (5).

$$y[i] = \sum_{k-1}^{K} x[i + d \cdot k] \cdot \omega[k] \tag{5}$$

In Equation (5), d denotes the dilation rate, k is the convolution kernel size, x represents the input feature map, and w indicates the convolution kernel weight. In the SE module, the squeezing equation of channel attention is shown in Equation (6).

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c} \tag{6}$$

In Equation (6), H and W respectively represent the height and width of the feature map, C is the number of channels, and Z_c is the channel statistical information obtained through global average pooling of spatial dimensions, which is used for subsequent excitation operations to generate channel weights and achieve enhancement of important features and suppression of redundant information. The study performs global average pooling on the feature map $x_{i,j,c}$ of each channel, compressing it into a 1×1×C vector to focus on the global information of the channel. The excitation equation of SE is shown in Equation (7).

$$s_c = \sigma\left(W_2 \cdot \mathrm{Re}\,LU\left(W_1 \cdot z_c\right)\right) \tag{7}$$

In Equation (7), W_1 and W_2 are the weight matrices for dimension reduction and dimension increase respectively, and σ represents the Sigmoid function, which outputs the weights from 0 to 1, representing the importance of this channel. Therefore, the modality data processing module composed of SE-ResNet-FCN and BERT after these improvements is shown in Figure 3.

As shown in Figure 3, the module preprocesses raw data for images and text. SE-ResNet optimizes image feature extraction in FCN, allowing the network to focus on key regions. Dilated convolution is applied in certain convolutional layers to capture global semantic information. The decoder then upsamples and fuses features, transforming low-resolution feature maps into high-resolution representations. BERT extracts semantic features from text and captures contextual dependencies. Finally, cross-modal fusion and alignment are performed to eliminate the modality gap between images and text, and the aligned features are output.

3.2 Multimodal retrieval model combining improved FCN and enhanced hash learning

Through SE-ResNet-FCN and BERT, the model extracts semantic features from images and texts, but this process alone does not complete cross-modal fusion. Hash learning addresses the efficiency and semantic association issues in multimodal retrieval by converting high-dimensional features of different modalities into low-dimensional binary hash codes. Because traditional Hash learning relies on manually designed features and is difficult to fully mine the complex semantic information of data, to solve this problem, the research adopts a dual-branch hash network structure. One branch is for images

and uses the image DH based on SE-ResNet-FCN, and the other branch is for text and uses the text DH based on BERT. Then, through the fusion layer, the features of different modalities are integrated and mapped to a unified hash code [18]. The dual-branch Hash network structure is shown in Figure 4.

As shown in Figure 4, the dual-branch Hash network maps semantically similar data from different modalities to nearby hash codes in the hash space, while pushing dissimilar data farther apart. After extracting features from the two branches and aligning their dimensions, the model fuses them through the fusion layer, then maps the unified feature to a binary hash code. A loss function constrains network training to ensure the semantic consistency of hash codes. The hash function mapping in Hash learning is shown in Equation (8)

$$h(x) = sign(f(x)) \tag{8}$$

In Equation (8), x represents the original high-dimensional feature, $f(x)$ is the hash mapping function, $sign(\cdot)$ is the sign function, and $h(x)$ represents the generated binary hash code. The mapping function in deep hashing is shown in Equation (9).

$$f(x) = W^T \cdot \phi(x) + b \tag{9}$$

In Equation (9), $\phi(x)$ represents the features extracted by the neural network, W and b are the learnable weight matrix and bias term. Hamming distance is used to measure sample similarity in the hash space. The equation for Hamming distance is shown in Equation (10).

$$d_H(h_a, h_b) = \frac{\frac{1}{k}\sum_{i=1}^{k}\left|h_a^{(i)} - h_b^{(i)}\right|}{2} \tag{10}$$

In Equation (10), h_a and h_b are two hash codes with length k, and $h_a^{(i)}$ and $h_b^{(i)}$ are the i-th bit of the hash code. Hamming similarity measures the similarity between two hash codes, as shown in Equation (11).

$$s_H(h_a, h_b) = 1 - d_H(h_a, h_b) \tag{11}$$

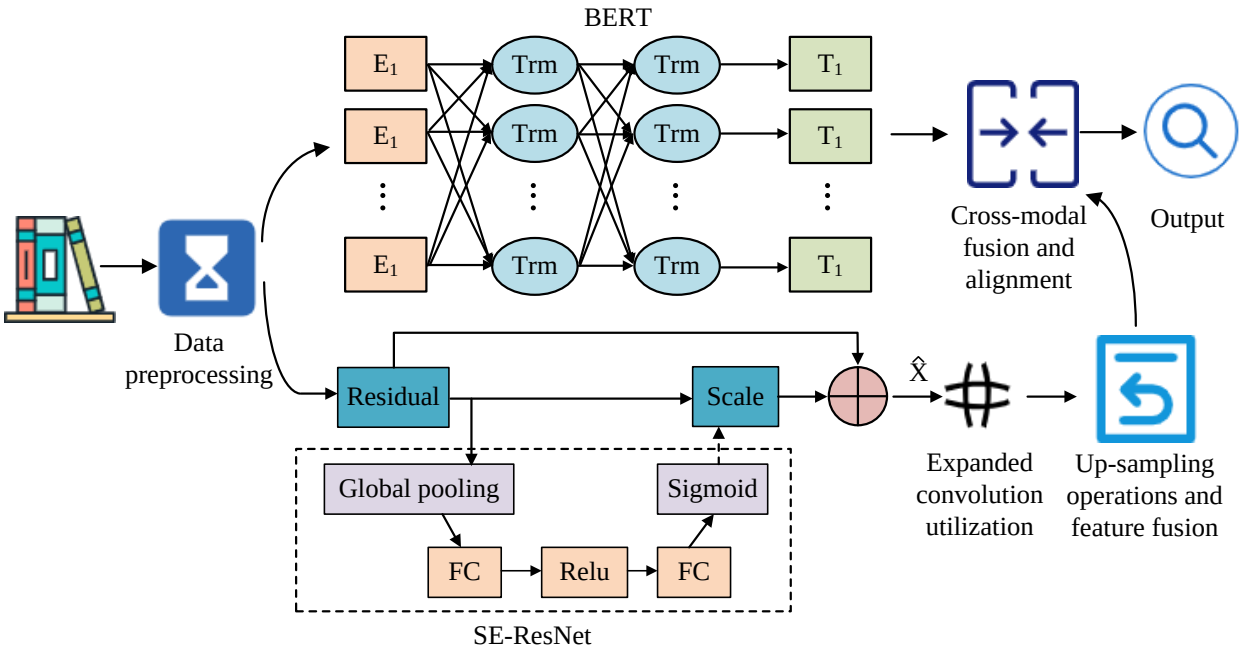


Figure 3: Modal data processing module composed of SE-ResNet-FCN and BERT.

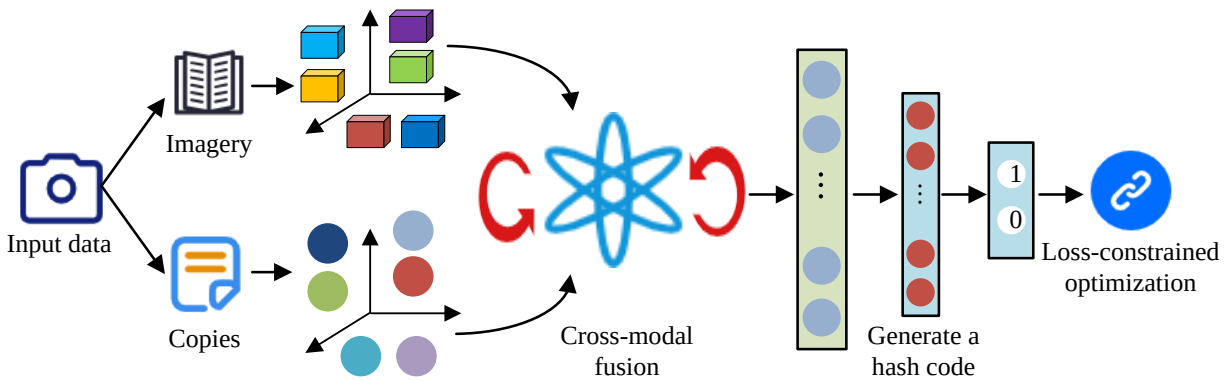


Figure 4: Double-branch Hash network structure diagram.

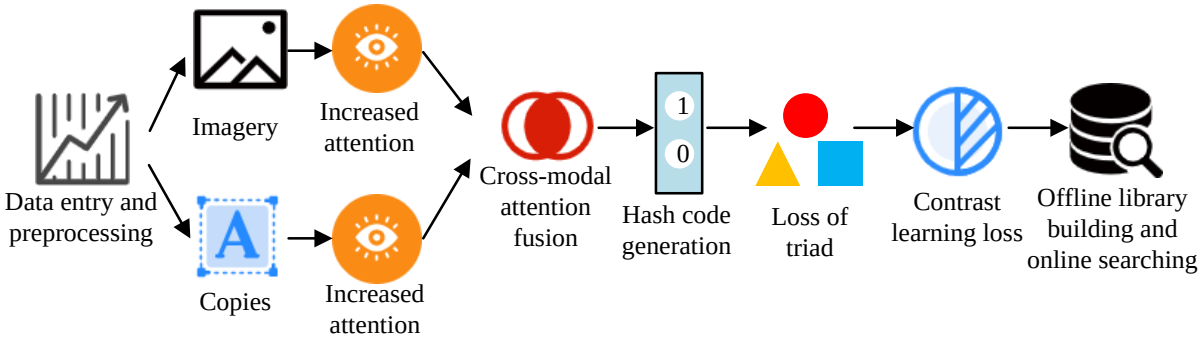


Figure 5: Improved dual-branch hash network structure diagram.

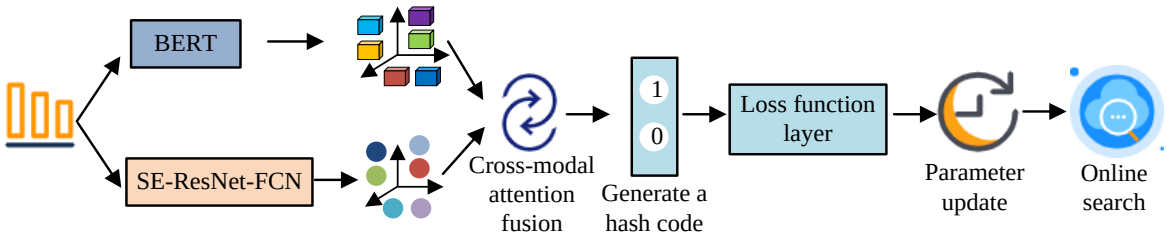


Figure 6: SRF-BERT-IDHS online library multimodal retrieval model.

In Equation (11), $s_H(h_a, h_b)$ represents the Hamming similarity, $d_H(h_a, h_b)$ is the length of the hash code, and the value range is [0,1]. Hamming similarity is complementary to Hamming distance. The dual-branch Hash network achieves cross-modal fusion to some extent, but the generated hash codes may only capture shallow patterns within each modality instead of deep semantic correlations. Triplet loss is introduced to enhance similarity constraints, reducing the distance between anchor and positive samples and increasing the distance to negative samples [19]. In addition, to improve the quality of hash codes, contrastive learning loss is introduced. By constructing positive and negative sample pairs, the similarity of positive sample pairs in the hash space is maximized, and the similarity of negative sample pairs is minimized. The structure diagram of the improved dual-branch hash network is shown in Figure 5.

As shown in Figure 5, the improved dual-branch Hash network focuses on hash code generation. Contrastive loss pulls the hash codes of similar samples closer, while triplet loss enforces the anchor-positive distance to be smaller than the anchor-negative distance. Through loss backpropagation, the network parameters are optimized so that the hash codes of semantically similar samples become closer, improving retrieval accuracy. The core equation of triplet loss is shown in Equation (12).

$$L_{triplet} = \max\left(d(h_a, h_p) - d(h_a, h_n) + \alpha, 0\right) \quad (12)$$

In Equation (12), $L_{triplet}$ is the triple loss function. h_a is the anchor sample hash code, h_p is the positive sample hash code with the same semantics, h_n is the negative sample hash code with different semantics, $d(\cdot)$ is the Hamming distance, and α is the margin parameter. The selection of marginal parameters for triplet loss is based on the distribution characteristics of positive and negative sample pairs during the training process, with a parameter adjustment range of 0.1 to 2.0. The contrastive learning loss is shown in Equation (13).

$$L_{contrast} = -\log \frac{\exp\left(s(h_v, h_t)/\tau\right)}{\sum_{k=1}^{N} \exp\left(s(h_v, h_{t,k})/\tau\right)} \quad (13)$$

In Equation (13), $L_{contrast}$ is the contrastive learning loss function. h_v represents the image hash code, h_t represents the matching text hash code, $h_{t,k}$ represents another text hash code, $s(\cdot)$ is the similarity function, and τ is the temperature parameter. In summary, the structure of the online library multimodal retrieval model (SRF-BERT-IDHS) composed of SRF-BERT and Improved Dual-branch Hash Structure (IDHS) modules is shown in Figure 6.

As shown in Figure 6, the unique novelty of the SRF-BERT-IDHS framework proposed in the study and its differences from previous multimodal hashing methods are as follows: Compared with the traditional dual-branch network structure, SRF-BERT-IDHS achieves dynamic alignment of image and text features at the hash mapping layer by introducing a cross-modal attention mechanism, effectively enhancing semantic consistency. In addition, by combining the joint optimization strategy of triple loss and contrastive learning, the compactness and discriminative ability of hash codes between heterogeneous modalities are further enhanced, overcoming the limitations of existing methods in semantic gaps and modal differences, thereby achieving efficient and accurate multimodal retrieval in large-scale online library scenarios.

Table 2: mAP test results at different hash code length levels.

Task	Model	Hash code length				
		16	32	64	128	256
I→T	SRF-BERT-IDHS	0.821	0.833	0.847	0.858	0.864
	LSSH	0.571	0.579	0.587	0.596	0.604
	CMFH	0.623	0.632	0.648	0.653	0.662
	SMFH	0.732	0.746	0.753	0.766	0.780
	DOCH	0.677	0.686	0.701	0.709	0.715
T→I	SRF-BERT-IDHS	0.872	0.889	0.905	0.916	0.931
	LSSH	0.678	0.698	0.705	0.712	0.726
	CMFH	0.712	0.723	0.734	0.745	0.751
	SMFH	0.783	0.798	0.810	0.823	0.831
	DOCH	0.724	0.735	0.747	0.760	0.775

The research proposes that the process of the model in real-time application or operation in large-scale library systems is divided into five stages: data upload, semantic feature extraction, hash coding, index storage and retrieval feedback. After the user inputs text or uploads an image, the system first calls the pre-trained SRF-BERT-IDHS model to extract the semantic features of the corresponding modal and generate a fixed-length hash code. Subsequently, the Hamming distance is calculated in the hash index database to quickly retrieve the closest cross-modal data. While returning the results, the system records the user's click behavior, duration of stay and feedback score as implicit feedback signals. Periodically, these feedback data are used to fine-tune the attention weights and hash mapping parameters of the model, optimizing the retrieval accuracy through an incremental learning mechanism. The learning rate is set to 0.001, the batch size is 64, and the Adam optimizer is used for parameter updates. The training lasted for 50 eras. The positive and negative samples in the triplet loss were sampled using the hard example mining strategy to ensure that the model focused on the discrimination near the semantic boundary during the optimization process. The training set and the test set are divided in an 8:2 ratio, and all data are randomly shuffled to ensure a balanced distribution of categories. The text encoder adopts the pre-trained BERT-base model, whose parameters are fixed in the early stage of training and then jointly updated at a smaller learning rate in the fine-tuning stage. The image backbone network is initialized based on the pre-trained SE-ResNet-FCN of ImageNet. The feature extraction process integrates the channel and spatial attention mechanism to enhance the response strength of key regions.

4 Validation of SRF-BERT-IDHS multimodal retrieval model for online libraries

4.1 Performance validation of SRF-BERT-IDHS model

To evaluate the multimodal retrieval performance of the SRF-BERT-IDHS model, it was compared with multimodal retrieval models for digital libraries based on Latent Semantic Sparse Hashing (LSSH), Collective Matrix Factorization Hashing (CMFH), Supervised Matrix Factorization Hashing (SMFH), and Discrete Online Cross-modal Hashing (DOCH). The experimental datasets were BookCover and ICDAR, both containing book image modalities and associated text information. The experiments were conducted on high-performance computing equipment with an Intel Core i9-10900K CPU (2.4 GHz, 32 cores) and 128 GB of memory. The software environment included Ubuntu 20.04 LTS and Python 3.7. The mean average precision (mAP) scores of the five models were compared under different hash code lengths, including image-to-text retrieval (I→T) and text-to-image retrieval (T→I) tasks. The performance comparison using the BookCover dataset is shown in Table 2.

In Table 2, the mAP value of SRF-BERT-IDHS in text-to-image retrieval reached a maximum of 0.931 and remained no lower than 0.872. SMFH ranked second, with a maximum mAP score of 0.831 and relatively stable performance. The mAP scores of all models in the text-to-image task were generally higher than those in the image-to-text task, indicating that text features enabled more precise image retrieval. Moreover, the hash code length was positively correlated with performance, suggesting that longer hash codes preserved more semantic information. To sum up, the SRF-BERT-IDHS model has the best performance. In all hash codes and two tasks, the mAP score of SRF-BERT-IDHS is higher than that of other models, and the advantage further expands with the increase of hash codes. To further verify the performance of SRF-BERT-IDHS, the five models were tested on both BookCover and ICDAR datasets, and their training efficiency was compared, as shown in Figure 7.

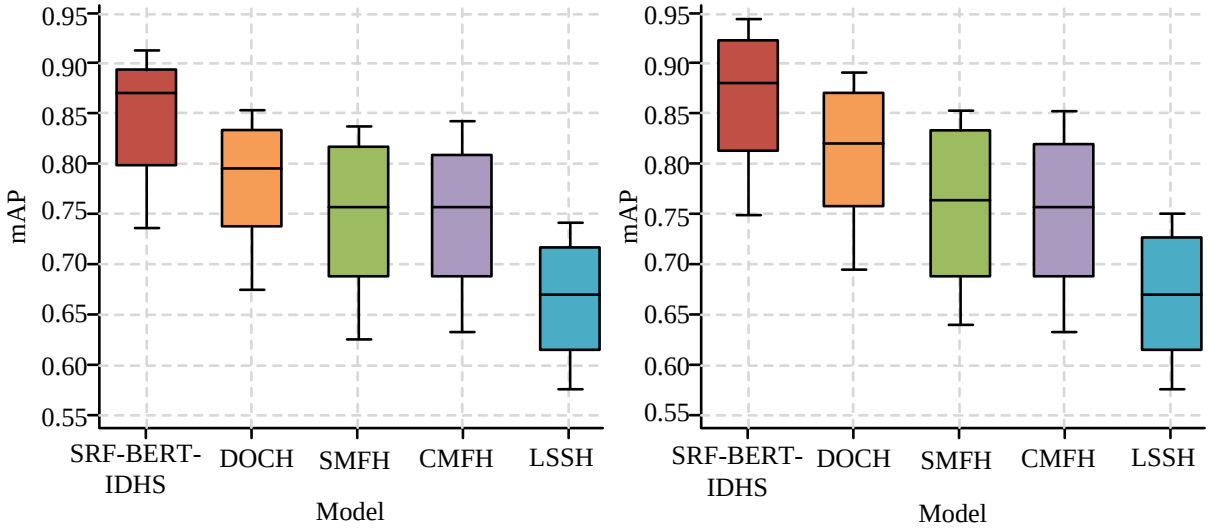
It can be seen from Figure 7 (a), the highest mAP score of SRF-BERT-IDHS was 0.92, the median was 0.87, and the lowest was 0.74. The lowest mAP score of LSSH was 0.58. As shown in Figure 7(b), the mAP scores of all models improved, with SRF-BERT-IDHS achieving a median mAP of 0.88 and a maximum of 0.95, which was 0.20 higher than the maximum of LSSH. Overall, SRF-BERT-IDHS demonstrated the highest precision and stability, ranking first among all compared models. This indicates that SRF-BERT-IDHS has strong generalization ability and robustness under different data distributions. Compared with other models, it can better adapt to the feature distribution of different datasets. Even in scenarios with complex semantics or significant cross-modal differences, it can still maintain high retrieval accuracy.

To comprehensively assess the learning capability of SRF-BERT-IDHS, the five models were trained on different datasets, and the comparison is shown in Figure 8.

As can be seen from Figure 8, all models exhibited slower loss reduction and higher overall loss on the BookCover dataset than on the ICDAR dataset, indicating that the relationship between book covers and text made the learning process more challenging. As shown in Figure 8(a), the loss of SRF-BERT-IDHS dropped sharply from $5.50E^{-09}$ to below $1.00E^{-09}$ within 20 epochs and was much lower than that of other models. The LSSH model consistently performed worst, with the highest loss reaching $2.00E^{-09}$. Compared with the ICDAR dataset, BookCover has higher requirements for the semantic alignment ability of the model, while SRF-BERT-IDHS can still converge rapidly on this dataset, indicating that it has stronger feature extraction and cross-modal matching capabilities. In contrast, the convergence speed of other models is relatively slow and they are prone to fall into local optimum, verifying the learning advantages of SRF-BERT-IDHS in complex scenarios.

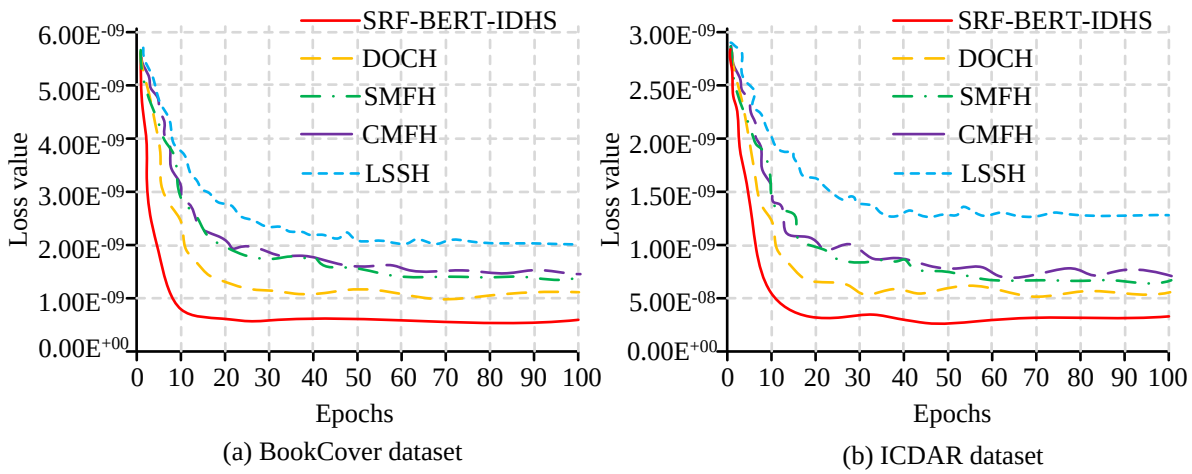
4.2 Practical application of multimodal retrieval model in online libraries

After verifying the basic performance of SRF-BERT-IDHS, its practical application value was further tested. A dataset of 5,000 pairs of book images (including cover and illustrations) and text information (including metadata, tables of contents, and main content) was collected from online digital libraries. The higher-quality 70% of the data was used as the training set, and the remaining 30% was used as the test set. The experimental environment remained unchanged. The five models were trained on the dataset, and their Average Precision (AP) was compared under different hash code lengths. The results are shown in Figure 9.



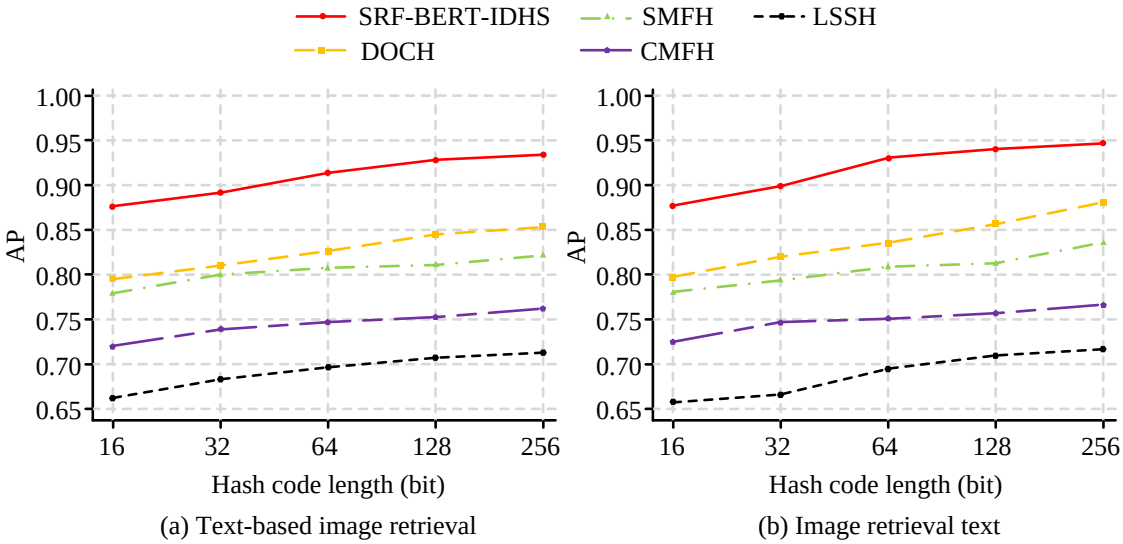
(a) BookCover dataset (b) ICDAR dataset

Figure 7: Training efficiency test results in different data sets.



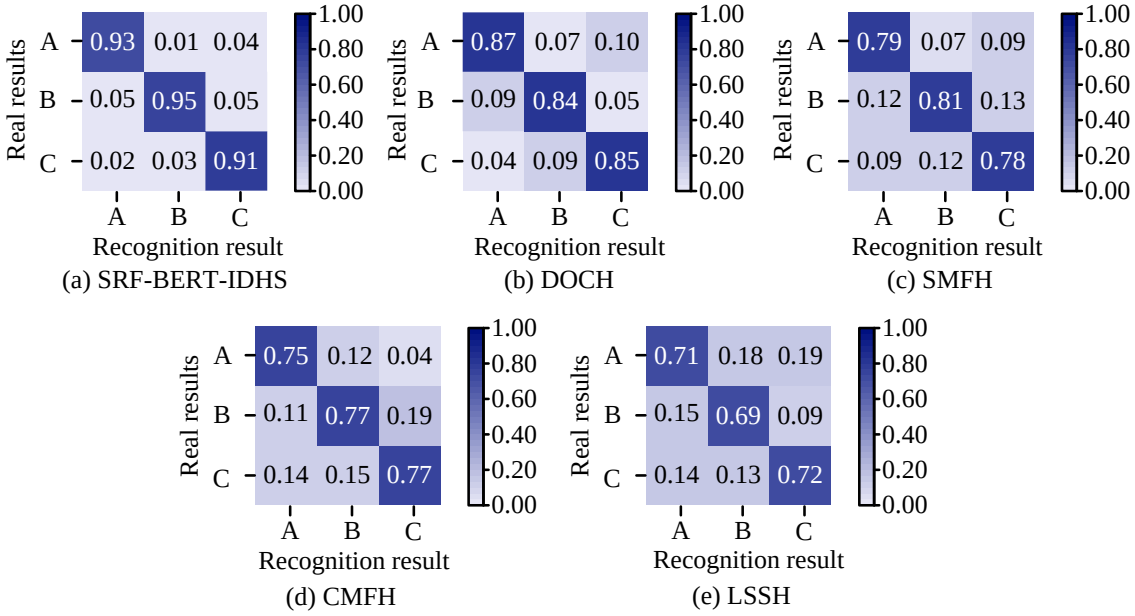
(a) BookCover dataset (b) ICDAR dataset

Figure 8: Comparison of learning capabilities of five models.



(a) Text-based image retrieval (b) Image retrieval text

Figure 9: Comparison of AP results in different hash codes.



(a) SRF-BERT-IDHS (b) DOCH (c) SMFH

(d) CMFH (e) LSSH

Figure 10: Comparison of retrieval confusion matrices of five models.

In Figure 9(a), the AP value of SRF-BERT-IDHS consistently outperformed the others, reaching 0.94 at 256-bit hash codes. The AP values of the other models also increased with longer hash codes, with DOCH reaching 0.85, which was 0.09 lower than SRF-BERT-IDHS. As shown in Figure 9(b), the AP value of SRF-BERT-IDHS further improved, reaching 0.95 at 256-bit hash codes. This is attributed to its integration of semantic association and feature reconstruction mechanisms, enabling more precise alignment of images and text in high-dimensional Spaces. On the 30% low-quality test set, SRF-BERT-IDHS still maintained an AP value of 0.91, significantly outperforming other models. The results show that this model has stronger robustness and generalization ability in real library retrieval scenarios. Especially when facing blurred images or incomplete texts, it can still maintain efficient matching performance, verifying its superior performance in practical applications. To evaluate the retrieval performance of SRF-BERT-IDHS on real datasets, the five models were tested on the training set, and their retrieval confusion matrices are shown in Figure 10.

It can be seen from Figure 10(a), SRF-BERT-IDHS achieved an average retrieval accuracy of 0.93 on the training set. As can be seen from Figure 10(b), DOCH performed slightly worse than SRF-BERT-IDHS, with an average accuracy of 0.85. From Figure 10(e), LSSH performed worst, with an average accuracy of only 0.71, which was 0.22 lower than SRF-BERT-IDHS. Overall, SRF-BERT-IDHS achieved the best recognition of multimodal data, with fewer misclassifications and the highest retrieval accuracy. To fully demonstrate the superior performance of SRF-BERT-IDHS, the top-N precision curves of the first 1,000 samples in the training set were compared across the five models, as shown in Figure 11.

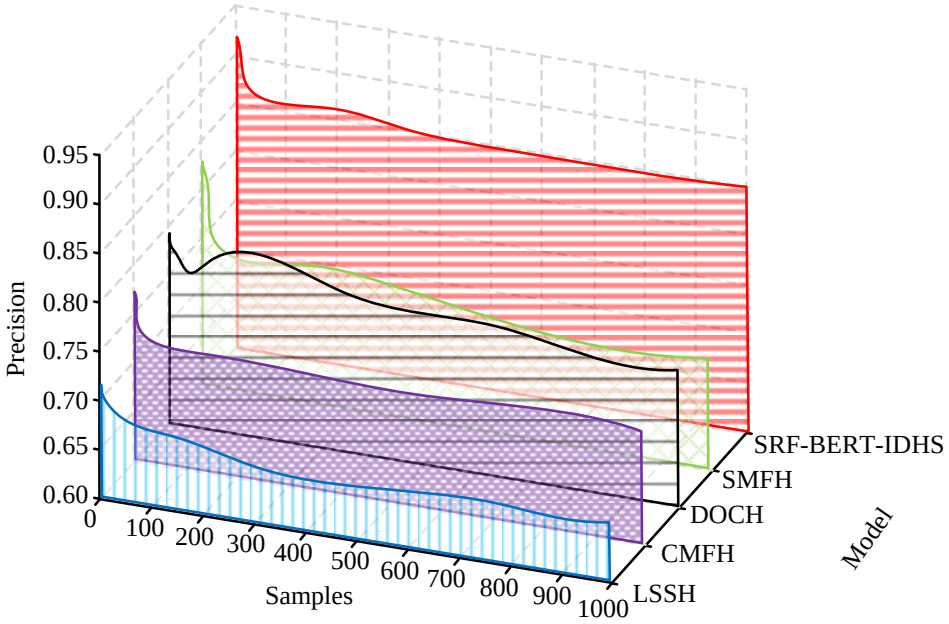


Figure 11: Comparison of accuracy curve results.

Table 3: Ablation study results on retrieval performance.

Project	Retrieval speed (s/ sample)	Training time (h)	Average accuracy rate (%)
FCN+SE-ResNet	0.034±0.02	12.5±1.2	76.8±2.2
FCN+Triplet loss only	0.029±0.03	11.8±1.6	79.3±2.4
FCN+Contrastive loss only	0.031±0.02	12.1±1.5	78.5±2.3
Full model	0.025±0.01*&#	10.3±1.3*&#	87.9±2.5*&#

Note: In Table 3, * indicates that the result difference between the Full model and FCN+SE-ResNet is significant (p<0.05); The results of the Full model and FCN+Triplet loss only represented by & were significantly different (p<0.05); # Indicates that the result difference between the Full model and FCN+Contrastive loss only is significant (p<0.05).

It can be seen from Figure 11, the top-N curve represents the variation of retrieval accuracy with the number of retrieval structure samples. The larger the area enclosed by the curve, the better the model performance. SRF-BERT-IDHS always maintains the final accuracy. As the sample size increases, the accuracy values of each model decrease. However, the decline of SRF-BERT-IDHS was the smallest, demonstrating stronger stability and generalization ability. Especially in the first 500 samples, its accuracy remained above 0.9, significantly superior to other models. The reason why SRF-BERT-IDH performs well is that it integrates semantic enhancement mechanisms and feature re-weighting strategies, effectively improving the alignment accuracy between text and image modalities. Meanwhile, the model introduces a context-aware semantic mapping module during the hash encoding process, significantly enhancing its robustness and retrieval stability in complex scenarios.

To test the effectiveness of the improved method proposed in the research, an ablation experiment was designed to analyze and compare its performance. The comparison indicators include retrieval speed, training time and average accuracy rate. The baselines of the comparative experiments include FCN+ SE-ResNet, FCN+Triplet loss only, FCN+Contrastive loss only, and Full model. The results are shown in Table 3.

As shown in Table 3, the complete model significantly outperforms each baseline method in terms of average accuracy, reaching 87.9%, and has the shortest training time, only 10.3 hours. This indicates that the introduced optimization strategy effectively enhances the model's convergence speed and retrieval accuracy. Compared with the variants that only use Triplet loss or Contrastive loss, the Full model further enhances cross-modal semantic consistency through the joint loss function and feature reweighting mechanism, verifying the effectiveness of the synergy of each module.

Although the BookCover dataset is representative in the task of book cover recognition, its sample distribution is limited to specific publication years and regional categories, making it difficult to comprehensively reflect the visual semantic differences across cultures and styles. To verify the generalization ability of the method, experiments were further conducted on the Flickr30K and MSCOCO datasets covering multi-domain image-text pairs. The results show that the proposed model maintains stable performance improvement under different semantic densities and noise levels, confirming its potential to adapt to diverse scenarios. Especially under complex backgrounds and low-quality image conditions, the model can still maintain a high retrieval accuracy rate, demonstrating good robustness.

5 Conclusion and future work

To address the bottleneck of multimodal retrieval in online libraries and to improve cross-modal retrieval efficiency and accuracy, SRF-BERT-IDHS explored a multimodal

retrieval approach for book images and text resources. The study built a retrieval framework that integrates FCN and Hash Learning. It used FCN to extract deep features of book cover images and combined them with text branch features. Through deep hash learning, the features were mapped to a low-dimensional semantic space. Triplet loss and contrastive learning loss were designed to optimize cross-modal semantic alignment. The results showed that the SRF-BERT-IDHS performed well in multimodal retrieval tasks. For the image-to-text retrieval task, the mAP reached 0.864 with a 256-bit hash code, which was higher than 0.604 for LSSH and 0.662 for CMFH. For the text-to-image retrieval task, the mAP reached 0.931 with a 256-bit hash code, with both precision and efficiency at a leading level. The AP value reached 0.940, and the average retrieval accuracy was 0.930. Even when the sample size reached 1000, the precision remained above 0.85. These results verified the capability of FCN to extract deep image features and the advantage of hash learning in dimension compression and semantic association enhancement. The combination effectively overcame modality barriers and met the multimodal resource retrieval needs of online libraries. However, although the dataset in The proposed method covered multiple book categories, its scale was relatively limited, and the robustness of the model under extreme long-tail distribution remained to be tested. When integrating other systems in the future, the research considers embedding the proposed model into the existing online library retrieval system, achieving efficient connection with the background database through API interfaces, and supporting real-time feature extraction and hash code matching. During the deployment process, a lightweight network structure and model compression technology are adopted. Meanwhile, an incremental learning mechanism is introduced to support online model updates and dynamic optimization, adapting to the new book entry and changes in user behavior.

6 Discussion

The cross-modal retrieval model of digital libraries proposed in the research maintains high accuracy and stable performance in both image and text retrieval. This is attributed to the deep characterization of image semantics by FCN and the effective modeling of cross-modal associations by hash coding. The deep shared proxy hash construction method mentioned in Reference [20] achieves a compact expression of cross-modal semantics through the proxy hash loss function, thereby enhancing the retrieval efficiency. This echoes the triplet loss and contrastive learning collaborative optimization strategy proposed in SRF-BERT-IDHS, both of which are dedicated to enhancing cross-modal semantic consistency. The federated cross-modal hashing method based on privacy enhancement culprits mentioned in reference [21] focuses on the balance between privacy protection and retrieval efficiency in distributed data storage. This method realizes cross-modal retrieval while ensuring user data privacy, providing a new idea for the distributed digital library scenario. Although SRF-BERT-IDHS did not directly involve privacy protection mechanisms, the constructed hash framework has good scalability. In the future, it can integrate federated learning strategies to enhance cross-modal retrieval capabilities while ensuring data security, further promoting the service upgrade of smart libraries and the development of trusted computing.

Pan R et al. proposed a knowledge base retrieval learning method to address the challenge of semantic consistent negation in image-text retrieval. This method enhances the accuracy and robustness of cross-modal semantic alignment by introducing a knowledge base and a lightweight cluster refinement strategy [22]. This is similar to the idea of using ResNet to replace the VGG network in FCN to enhance the ability of image feature extraction, both aiming to strengthen the semantic consistency between modalities. The residual structure of ResNet effectively alleviates the degradation problem of deep networks, enabling the model to have stronger representational capabilities when processing complex images and thereby enhancing the accuracy of cross-modal matching. The research method poses certain challenges in future practical applications and expansions. It is necessary to consider the balance between the computing resource consumption of model deployment and the timeliness of response. Especially in scenarios with large-scale concurrent user access, the efficiency of hash code generation and the design of index structure need to be further optimized. The efficient cross-modal feature matching model based on the CLIP framework mentioned in Reference [23] divides the model into two parts: feature extraction and contrastive learning. By pre-training a large model, it realizes the unified semantic space mapping of images and text, significantly improving the cross-modal matching efficiency. This idea provides an important reference for optimizing the feature extraction module in SRF-BERT-IDHS. In the future, it can be combined with lightweight CLIP variants to reduce computational overhead while maintaining high-precision retrieval performance.

Based on the above content, it can be known that the cross-modal retrieval method proposed in the research shows significant advantages in semantic alignment accuracy and model scalability. Especially after combining the triple loss and contrastive learning mechanism, the association expression ability between images and texts is further enhanced. In the future, practical applications and deployments need to be oriented towards the complexity of real scenarios. It is advisable to consider introducing a dynamic adaptive hash code length adjustment mechanism to address the matching deviation caused by the distribution differences of different modal data. Meanwhile, in combination with the edge computing architecture, some feature extraction tasks are decentralized to terminal devices to reduce the load pressure on the central server.

References

[1] Shubhi Bansal, Mohit Kumar, Chandravardhan Singh Raghaw, and Nagendra Kumar. Sentiment and hashtag-aware attentive deep neural network for

multimodal post popularity prediction. Neural Computing and Applications, 37(4):2799-2824, 2025. https://doi.org/10.1007/s00521-024-10755-5

[2] Yangdong Chen, Jiaqi Quan, Yuejie Zhang, Rui Feng, and Tao Zhang. Deep cross-modal hashing with fine-grained similarity. Applied Intelligence, 53(23):28954-28973, 2023. https://doi.org/10.1007/s10489-023-05028-y

[3] Prabhjot Kaur, and Chander Kant. 2-Phase multi-trait biometric authentication model against spoofing attack using deep hash model. SN Computer Science, 6(1):47-59, 2025. https://doi.org/10.1007/s42979-024-03513-w

[4] Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. Learning explicit and implicit dual common subspaces for audio-visual cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 19(2s):1-23, 2023. https://doi.org/10.1145/3564608

[5] Jiaxing Li, Wai Keung Wong, Lin Jiang, Xiaozhao Fang, Shengli Xie, and Yong Xu. CKDH: CLIP-based knowledge distillation hashing for cross-modal retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 34(7):6530-6541, 2024. https://doi.org/10.1109/TCSVT.2024.3350695

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 139:8748-8763, 2021. https://doi.org/10.48550/arXiv.2103.00020

[7] Jie Lin, Olivier Morere, Vijay Chandrasekhar, Antoine Veillard, and Hanlin Goh. DeepHash: Getting regularization, depth and fine-tuning right. ACM Transactions on Multimedia Computing, Communications and Applications, 18(1):1-21, 2022. https://doi.org/10.48550/arXiv.1501.04711

[8] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1):539-546, 2018. https://doi.org/10.1609/aaai.v32i1.11263

[9] Zhenqiu Shu, Li Li, Jun Yu, Donglin Zhang, Zhengtao Yu, and Xiao-Jun Wu. Online supervised collective matrix factorization hashing for cross-modal retrieval. Applied Intelligence, 53(11):14201-14218, 2023. https://doi.org/10.1007/s10489-022-04189-6

[10] Saeid Sattari, Sinan Kalkan, and Adnan Yazici. Multimodal multimedia information retrieval through the integration of fuzzy clustering, OWA-based fusion, and Siamese neural networks. Fuzzy Sets and Systems, 515:109419, 2025. https://doi.org/10.1016/j.fss.2025.109419

[11] Donglin Zhang, Xiao-Jun Wu, and Guoqing Chen. ONION: Online semantic autoencoder hashing for cross-modal retrieval. ACM Transactions on Intelligent Systems and Technology, 14(2):1-18, 2023. https://doi.org/10.1145/3572032

[12] Asad Khan, Sakander Hayat, Muhammad Ahmad, Jinyu Wen, Muhammad Umar Farooq, Meie Fang, and Wenchao Jiang. Cross‐modal retrieval based on deep regularized hashing constraints. International Journal of Intelligent Systems, 37(9):6508-6530, 2022. https://doi.org/10.1002/int.22853

[13] Yongxin Wang, Yu-Wei Zhan, Zhen-Duo Chen, Xin Luo, and Xin-Shun Xu. Multiple information embedded hashing for large-scale cross-modal retrieval. IEEE Trans Circuits Syst Video Technol, 34(6):5118-5131. https://doi.org/10.1109/TCSVT.2023.3340102

[14] Haokun Wen, Xuemeng Song, Jianhua Yin, Jianlong Wu, Weili Guan, and Liqiang Nie. Self-training boosted multi-factor matching network for composed image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5):3665-3678, 2023. https://doi.org/10.1109/TPAMI.2023.3346434

[15] Xingwei Zhang, Xiaolong Zheng, Wenji Mao, and Daniel Dajun Zeng. Boosting deep cross-modal retrieval hashing with adversarially robust training. Applied Intelligence, 53(20):23698-23710, 2023. https://doi.org/10.1007/s10489-023-04715-0

[16] Yu Liu, Haipeng Chen, Guihe Qin, Jincai Song, and Xun Yang. Bias mitigation and representation optimization for noise-robust cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 21(11):1-17, 2025. https://doi.org/10.1145/3700596

[17] Yuxia Cao. Joint feature fusion hashing for cross-modal retrieval. International Journal of Machine Learning and Cybernetics, 15(12):6149-6162, 2024. https://doi.org/10.1007/s13042-024-02309-x

[18] Venice Erin Liong, Jiwen Lu, Gang Wang, Pierre Moulin, and Jie Zhou. Deep hashing for compact binary codes learning. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2475-2483, 2015. https://doi.org/10.1109/cvpr.2015.7298862

[19] Ying Ma, Meng Wang, Guangyun Lu, and Yajun Sun. Multi-label semantic sharing based on graph convolutional network for image-to-text retrieval. The Visual Computer, 41(3):1827-1840, 2025. https://doi.org/10.1007/s00371-024-03496-y

[20] Lirong Han, Mercedes E. Paoletti, Sergio Moreno-Álvarez, Juan M. Haut, and Antonio Plaza. Deep shared proxy construction hashing for cross-modal remote sensing image fast target retrieval. ISPRS Journal of Photogrammetry and Remote Sensing, 218(1):44-56, 2024. https://doi.org/10.1016/j.isprsjprs.2024.10.004

[21] Ruifan Zuo, Chaoqun Zheng, Fengling Li, Lei Zhu, and Zheng Zhang. Privacy-enhanced prototype-based federated cross-modal hashing for cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 20(9):1-19, 2024. https://doi.org/10.1145/3674507

[22] Renjie Pan, Hua Yang, and Xiangyu Zhao. Real: Improving image-text retrieval with authentic negative repository learning. ACM Transactions on Multimedia Computing, Communications and Applications, 21(10):1-22, 2025. https://doi.org/10.1145/3729172

[23] Yilin Peng. A CLIP-based cross-modal matching model for image-text retrieval. Information Technology and Control, 54(3):1030-1048, 2025. https://doi.org/10.5755/j01.itc.54.3.41801